Agnostic risk parity: taming known and unknown unknowns
Raphael Benichou, Yves Lempérière, Emmanuel Sérié, Julien Kockelkoren, Philip Seager, Jean-Philippe Bouchaud and Marc Potters

Black–Litterman, exotic beta and varying efficient portfolios: an integrated approach
Ricky Alyn Cooper and Marat Molyboga

Correctness of backtest engines
Robert Löw, Stanislaus Maier-Paape and Andreas Platen

Statistical testing of DeMark technical indicators on commodity futures
Marco Lissandrin, Donnacha Daly and Didier Sornette
Risk.net in numbers

- **200+** New articles & technical papers (each month)
- **140,000** Users (each month)
- **370,000** Page views (each month)

58,000+ articles stretching back 20 years

- **6,500+** on Asset Management
- **6,900+** on Commodities
- **£21,000+** on Derivatives
- **19,400+** on Regulation
- **19,600+** on Risk Management

See what you’re missing

Visit the world’s leading source of exclusive in-depth news & analysis on risk management, derivatives and complex finance now.

To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
The Journal of Investment Strategies
EDITORIAL BOARD

Editor-in-Chief
ARThUR M. BERD General Quantitative LLC

Advisory Board
ROBERT ENGLE NYU Stern School of Business
KENNETH A. FROOT Harvard Business School
ROBERT JARROW Cornell University, Johnson School of Business

Editorial Board
JESPER ANDREASEN Danske Bank
MARCO AVELLANEDA NYU Courant Institute
VINEER BHANSALI PIMCO
JEAN-PHILIPPE BOUCHAUD Capital Fund Management
PETER CARR Morgan Stanley
JOHN CHISHOLM Acadian Asset Management
DEAN CURNUTT Macro Risk Advisors
VLADIMIR FINKELSTEIN Blue Mountain Capital
ROSS GARON Cubist Systematic Strategies
LISA GOLDBERG UC Berkeley
PHILIPPE GOUGENHEIM Gougenheim Investments
ALI HIRSA Sauma Capital LLC
PHILIPPE ITHURBIDE Amundi Asset Management

SUBSCRIPTIONS

The Journal of Investment Strategies (Print ISSN 2047-1238 | Online ISSN 2047-1246) is published quarterly by Infopro Digital Risk Limited, Haymarket House, 28–29 Haymarket, London SW1Y 4RX, UK. Subscribers

Subscription packages are available on an annual basis. To find out about the different options, including our exclusive academic rates starting from £100, visit subscriptions.risk.net/journals-print or contact ben.anderton@infopro-digital.com.

All prices include postage. All subscription orders, single/back issues orders, and changes of address should be sent to:

UK & Europe Office: Infopro Digital, Haymarket House, 28-29 Haymarket, London SW1Y 4RX, UK. Tel: +44 1858 438 800

US & Canada Office: Infopro Digital, 55 Broad Street, Floor 22, New York, NY, 10005, USA. Tel: +1 212 776 8075

Asia & Pacific Office: Infopro Digital, Unit 1704–05, Berkshire House, Taikoo Place, 25, Westlands Road, Quarry Bay, Hong Kong. Tel +852 3411 4828

Website: risk.net/journals
E-mail: info@risk.net

To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
The Journal of Investment Strategies
GENERAL SUBMISSION GUIDELINES

Manuscripts and research papers submitted for consideration must be original work that is not simultaneously under review for publication in another journal or other publication outlets. All articles submitted for consideration should follow strict academic standards in both theoretical content and empirical results. Articles should be of interest to a broad audience of sophisticated practitioners and academics.

Submitted papers should follow Webster’s New Collegiate Dictionary for spelling, and The Chicago Manual of Style for punctuation and other points of style, apart from a few minor exceptions that can be found at www.risk.net/journals. Papers should be submitted electronically via email to: journals@infopro-digital.com. Please clearly indicate which journal you are submitting to.

You must submit two versions of your paper; a single \LaTeX{} version and a PDF file. \LaTeX{} files need to have an explicitly coded bibliography included. All files must be clearly named and saved by author name and date of submission. All figures and tables must be included in the main PDF document and also submitted as separate editable files and be clearly numbered.

All papers should include a title page as a separate document, and the full names, affiliations and email addresses of all authors should be included. A concise and factual abstract of between 150 and 200 words is required and it should be included in the main document. Five or six keywords should be included after the abstract. Submitted papers must also include an Acknowledgements section and a Declaration of Interest section. Authors should declare any funding for the article or conflicts of interest. Citations in the text must be written as (John 1999; Paul 2003; Peter and Paul 2000) or (John et al 1993; Peter 2000).

The number of figures and tables included in a paper should be kept to a minimum. Figures and tables must be included in the main PDF document and also submitted as separate individual editable files. Figures will appear in color online, but will be printed in black and white. Footnotes should be used sparingly. If footnotes are required then these should be included at the end of the page and should be no more than two sentences. Appendixes will be published online as supplementary material.

Before submitting a paper, authors should consult the full author guidelines at:

http://www.risk.net/static/risk-journals-submission-guidelines

Queries may also be sent to:

The Journal of Investment Strategies, Infopro Digital,
Haymarket House, 28–29 Haymarket, London SW1Y 4RX, UK
Tel: +44 1858 438 800 (UK/ EU), +1 212 776 8075 (USA),
+852 3411 4828 (Asia)
E-mail: Journals@infopro-digital.com

To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
The journal

*The Journal of Investment Strategies* is an international refereed journal focusing on the rigorous treatment of modern investment strategies.

Investment strategy research is a distinct subject matter, combining insights from various areas of financial economics and utilizing techniques from econometrics and statistics, among other fields. As an applied field of research, it directly impacts the practice of asset management, a large and diverse industry with many constituents, including traditional and alternative buy-side investment managers as well as the sell-side and independent advisors. As an academic topic, it presents unique and interesting challenges for understanding the sources of investment returns and for formulating consistent and systematic methodologies for portfolio management in a dynamic context.

Much research in this area, particularly the studies going beyond classical portfolio theory, is not currently accessible to a wider audience. In particular, many research papers originating from the industry are not well distributed. *The Journal of Investment Strategies*, therefore, has three fundamental aims:

1. to foster high-quality, original and innovative work on investment strategies;
2. to provide practitioners and academics with access to the resulting technical research; and
3. to serve as an educational forum on timely issues concerning investment strategies.

**Content Guidelines**

Topics considered for publication in the journal include:

- Fundamental strategies;
- Relative value strategies;

To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
● Tactical strategies;
● Event-driven strategies;
● Algorithmic trading strategies;
● Principal investment strategies;
● Portfolio management and asset allocation;
● Econometric and statistical methods.

The Editorial Board will consider two types of submission:

● Full-length original papers covering both theory and practice; and
● Short expository or discussion papers for publication in the Investment Strategy Forum.

Research Papers
Research papers are typically full-length academic papers of between 8,000 and 10,000 words in length. Topics for research can be of interest to both academic and industry professionals. Papers submitted for publication must be original and should not have been published or considered for publication elsewhere.

Investment Strategy Forum
The Forum is intended to provide rapid communication on findings and ideas about investment strategies that are topical, expository and educational in nature. The mission of the Forum is to promote active discussions of current issues. Articles should not exceed 6,000 words. The main goal of these submitted papers is to educate a wider audience and to increase understanding of the issues and topics that may not be easily accessible to either academics or practitioners.
CONTENTS

Letter from the Editor-in-Chief vii

RESEARCH PAPERS

Agnostic risk parity: taming known and unknown unknowns 1
Raphael Benichou, Yves Lempérière, Emmanuel Sérié, Julien Kockelkoren, Philip Seager, Jean-Philippe Bouchaud and Marc Potters

Black–Litterman, exotic beta and varying efficient portfolios: an integrated approach 13
Ricky Alyn Cooper and Marat Molyboga

Correctness of backtest engines 31
Robert Löw, Stanislaus Maier-Paape and Andreas Platen

Statistical testing of DeMark technical indicators on commodity futures 53
Marco Lissandrin, Donnacha Daly and Didier Sornette

Editor-in-Chief: Arthur M. Berd Subscription Sales Manager: Ben Anderton
Publisher: Nick Carver Global Key Account Sales Director: Michelle Godwin
Journals Manager: Sarah Campbell Composition and copyediting: T&T Productions Ltd
Printed in UK by Printondemand-Worldwide

©Copyright Infopro Digital Risk (IP) Limited, 2017. All rights reserved. No parts of this publication may be reproduced, stored in or introduced into any retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the copyright owners.
LETTER FROM THE EDITOR-IN-CHIEF

Arthur M. Berd
Founder and CEO, General Quantitative LLC

Welcome to the third issue of the sixth volume of The Journal of Investment Strategies. We present here two different approaches to the perennial problem of robust portfolio construction, a deep examination of backtesting engines, and a rigorous take on a long-standing technical trading technique.

In the issue’s first paper, “Agnostic risk parity: taming known and unknown unknowns”, Raphael Benichou, Yves Lempérière, Emmanuel Sérié, Julien Kockelkoren, Philip Seager, Jean-Philippe Boucaud and Marc Potters – the research team at Capital Fund Management (CFM) – present a novel take on the notion of risk parity. They identify their paper’s main goal as the construction of a portfolio that will result in a robust out-of-sample diversification: in other words, a portfolio that truly is equally driven by all of the relevant driving factors. Unlike the more traditional approaches to risk parity, the CFM team focuses on the notion of managing not only measurable and forecastable risks (known unknowns) but also unmeasurable and unforecastable uncertainties (unknown unknowns) when building resilient portfolios. In a signature physicist style, they solve the problem by positing a deeper symmetry requirement – a rotationally invariant symmetry of the portfolio representation through assets – and derive the plausible solution from it with minimal additional assumptions.

This symmetry requirement has a solid reason behind it: without it, we may be able to achieve good diversification with regard to one definition of “fundamental assets” while failing to achieve any reasonable diversification under another definition of fundamental assets. The notion that we might need to think more deeply about the definition of assets is not an idle thought, given that in many cases portfolio managers actually implement their views with derivatives, which can be easily redefined to include or exclude any particular exposure. The results of this paper will likely be most interesting to quants running actively traded portfolios, such as diversified futures trend-following funds, which have recently seen a challenging period in terms of performance. A better portfolio construction methodology will be a welcome development for these investors.

“Black–Litterman, exotic beta and varying efficient portfolios: an integrated approach” is the second paper in this issue. Here, Ricky Alyn Cooper and Marat Molyboga investigate the interrelationships among the Black–Litterman, exotic beta and risk parity approaches to robust portfolio construction. The authors show that these approaches need not be considered as alternatives to each other; instead, they can be considered as being part of an integrated framework, unifying the various insights. In particular, the authors recast the exotic beta portfolios as “views” within
the Black–Litterman framework. This part is not controversial at all, as exotic betas represent potential robust outperformance sources, and therefore the exposures to such betas should indeed be consistent with the Black–Litterman approach, in which a portfolio manager is free to express their forecasts with respect to the relative outperformance of certain positions compared with equilibrium returns.

The second contention of Cooper and Molyboga – that one can use risk parity portfolios instead of equilibrium (market cap) ones as a starting point for Black–Litterman methodology – is both more interesting and more controversial. It is interesting because it is quite possible that, as the authors show, such a starting point leads to potentially better end results and better stability. However, it is also controversial because we no longer have as firm a logical foundation as in the classical interpretation. A leap of faith is made to assume that a portfolio with a higher Sharpe ratio, such as a risk parity portfolio, may in fact be acceptable as the Bayesian prior in the Black–Litterman method.

I do not have any practical objections to this claim: it is entirely plausible. However, whether the entire world would consider this an incontrovertible truth remains to be seen and, in fact, if they did, it would probably have such an impact on the market that the risk parity portfolio would become identical to the market cap, i.e., it would clear the supply and demand.

Robert Löw, Stanislaus Maier-Paape and Andreas Platen investigate a tricky concept in “Correctness of backtest engines”, our third paper. As most practitioners of investment management, along with many academics, rely on such engines, either commercial or homegrown, for conducting their research, the results of this paper and the authors’ cautionary words should be of great importance for a large audience. The main problem the authors point out is to do with execution assumptions under incomplete information, such as when using compressed open–low–high–close (OLHC) bars and even tick data. They caution against believing the results that assume the strategy algorithm could have actually traded at a trigger price at the time when such a price triggers an order in our strategy. They offer several unit tests that can be incorporated into algorithmic trading software to guard against such creeping inconsistencies, which would then inevitably lead to bad out-of-sample performance. Correspondingly, I believe the issues highlighted by the authors are very important, and that anyone using backtest software should check whether their results are, in fact, affected by these problems.

While I concur with the authors’ focus on the consistency of backtest software, I think these important problems can be mitigated in most settings by following a carefully and conservatively designed algorithmic strategy. For example, if the strategy, as a general rule, includes a time lapse between the decision making and the emission of the trade order, then such a strategy will be more robust to the issues highlighted by the authors. Some more advanced backtesting software vendors actually include this
time lapse feature in their software settings, so the researcher does not even need to explicitly account for it in the strategy: it is already taken care of within the backtest module.

In the fourth paper, “Statistical testing of DeMark technical indicators on commodity futures”, Marco Lissandrin, Donnacha Daly and Didier Sornette undertake a deep study of a few of the best-known technical trading indicators: the DeMark Sequential, Combo and Setup trade indicators. It is admittedly uncommon to see a piece of deep, quantitative research covering the decidedly nonacademic topic of technical trading. However, since these indicators are actually quite popular, with many individual and even professional traders relying on them and often just following the output of some software or data subscription vendor, we think a deeper dive into this subject is well justified. The authors construct sophisticated random resampling tests to verify the statistical significance of the indicators. Further, they show that, in the case of commodity futures trading, the manner in which the futures roll is implemented affects the expected performance of the technical trading. With such in-depth analysis, this paper will certainly be a necessary read for many strategists utilizing technical trading methods.

In conclusion, I hope this issue of The Journal of Investment Strategies will find a broad audience, and that readers will not only read the paper that first catches their eye but also look through the other papers in the issue, as such a cross-pollination of ideas is what drives creativity and innovation in the field. It was certainly one of the original objectives of our journal, and I am pleased that we have been able to continue in this vein for over six years.
Research Paper

Agnostic risk parity: taming known and unknown unknowns

Raphael Benichou, Yves Lempérière, Emmanuel Sérié, Julien Kockelkoren, Philip Seager, Jean-Philippe Bouchaud and Marc Potters

Capital Fund Management, 23 Rue de l'Université, 75007 Paris, France; emails: raphael.benichou@cfm.fr, yves.lempiere@cfm.fr, emmanuel.serie@cfm.fr, julien.kockelkoren@cfm.fr, philip.seager@cfm.fr, jean-philippe.bouchaud@cfm.fr, marc.potters@cfm.fr

(Received October 31, 2016; accepted January 16, 2017)

ABSTRACT

Markowitz’s celebrated optimal portfolio theory generally fails to deliver out-of-sample diversification. In this paper, we propose a new portfolio construction strategy based on symmetry arguments only, leading to “eigenrisk parity” portfolios that achieve equal realized risk on all the principal components of the covariance matrix. This holds true for any other definition of uncorrelated factors. We then specialize our general formula to the most agnostic case, where the indicators of future returns are assumed to be uncorrelated and of equal variance. This “agnostic risk parity” (AGP) portfolio minimizes unknown-unknown risks generated by the over-optimistic hedging of different bets. AGP is shown to fare quite well when applied to standard technical strategies such as trend following.

Keywords: quantitative portfolio management; portfolio construction; risk parity; risk budgeting; diversification; Markowitz.
1 INTRODUCTION

Diversification is the mantra of rational investment strategies. Harry Markowitz proposed a mathematical incarnation of that mantra that is common lore in the professional world. Unfortunately, the practical implementation of Markowitz’s ideas is fraught with difficulties and yields very disappointing results. This has been known for a long time, with many papers attempting to identify its flaws and suggesting remedies (Black and Litterman 1992; Choueifaty and Coignard 2008; Deguest et al 2013; Meucci 2009; Michaud 1998; Roncalli 2013). The most important problems are well understood: the optimally diversified Markowitz portfolio, somewhat paradoxically, often ends up being very concentrated on a few assets only, which inevitably leads to disastrous out-of-sample risks. The optimal portfolio is also unstable in time and sensitive to small changes in parameters and/or expected future gains. In the face of these difficulties, two distinct branches of research have emerged.

The first branch concerns the determination of the covariance matrix of the $N$ different assets eligible in the portfolio, eg, all the stocks belonging to a given index. This covariance matrix is specified by a large number of entries $(N \times (N + 1)/2)$ for which only a limited amount of data is available $(N \times T$, where $T$ is the length of the time series at one’s disposal). When $T$ is not extremely large compared with $N$, the empirically determined covariance matrix is highly unreliable and leads to severe instabilities when used in the Markowitz optimization program. Recently, some powerful mathematical tools have been proposed to optimally “clean” the empirical covariance matrix, leading to a very significant improvement in the efficiency of Markowitz diversification using the so-called rotationally invariant estimator (RIE). For a short review of this, see Bun et al (2017) and the references therein.

Another crucial step, of course, is to specify a list of expected returns for each asset. These expected returns result from either quantitative signals (such as trend following) or other forms of analysis (quantitative or subjective). These signals are usually extremely noisy and unreliable, so one should instead speak, as we shall do below, of indicators, ie, possibly suboptimal and biased predictions of future returns.

Once all this is done, however, a timeworn but fundamental problem remains (Keynes 1934; Taleb 2010). Even when sophisticated statistical tools can adequately deal with risk, they cannot handle uncertainty, ie, the intrinsic propensity of financial markets to behave in a way that is not consistent with prior probabilities. For example, although the future “true” covariance matrix is often reasonably close to the cleaned (RIE) covariance matrix, correlations can suddenly shift to a new regime that was never observed in the past. This is worse for expected returns that are even more exposed to unknown unknowns than volatility or correlations. One therefore needs an extra layer of control, beyond Markowitz’s optimization, that acts as a safeguard against statistically unexpected events.
This is what the second strand of research mentioned above attempts to address. The idea is to add to the standard risk–return objective function some extra penalty terms that enforce diversification, typically in the form of generalized Herfindahl indexes or entropy functions (Bouchaud et al. 1997; Frahm and Wiechers 2011; Meucci 2009). This has led to important breakthroughs, such as the concept of maximally diversified portfolios (MDP; Choueifaty and Coignard (2008)) or, more recently, principal risk parity portfolios (PRP), with several variations on this theme (see Bailey and Lopez de Prado 2012; Deguest et al. 2013; Kind 2012; Lohre et al. 2012; Partovi and Caputo 2004).

2 DIVERSIFICATION AND ISOTROPY

Although interesting, there is a hidden assumption in these penalty terms that is far from neutral; this is the choice of assets one considers as “fundamental”, among which risk should be as diversified as possible in the portfolio. These assets are chosen to be physical stocks for MDPs, or the principal components of the correlation matrix in the case of PRPs. In the case of long-only portfolios and traditional asset management, the choice of physical assets as the natural “basis” for portfolio construction might be reasonable. But for, say, a portfolio of futures contracts with long and short positions, any linear combination of these assets is a priori feasible (at least within some overall leverage constraint). In mathematical terms, one can “rotate” the natural asset basis into any a priori equivalent one. The point, however, is that a MDP in one basis can, in fact, become maximally concentrated in another. Take, for example, a portfolio of stocks with equal weights $w_i = 1/N$ on all $N$ stocks. From the point of view of the (neg)entropy $S = \sum_i w_i \ln w_i$, or of the Herfindahl index $H = \sum_i w_i^2$, this is clearly optimal. However, since the leading risk factor associated with the correlation matrix is itself very close to an equi-weighted allocation on all stocks, a rotation onto the principal component basis $\alpha$ leads to the worst possible values for both the entropy and the Herfindahl index. In other words, the very concept of maximal diversification is not invariant under a redefinition of the assets considered as “fundamental”. Another vivid example of the arbitrariness in the definition of fundamental assets is provided by the interest rate curve, or, more generally, by contracts with different maturities. Should one consider the physical contracts, or only one of them and all associated calendar spreads?

Are there directions in asset space that play a special role? Can one unambiguously elicit risk factors that are more fundamental than others? This is an old problem in quantitative finance, and there is a long list of papers that have attempted to identify these factors, particularly in the equity space. However, as recently reviewed by Puthauanthurong and Roll (2014), there is no consensus on this point. If risk is associated
with volatility (or variance), then the problem is completely degenerate, or, using mathematical parlance, isotropic.

To make this clear, let us consider asset returns \( r_i \) \((i = 1, \ldots, N)\) as random variables, with zero mean and a (true) covariance matrix \( C \), where \( C_{ij} = \mathbb{E}[r_i r_j] \).

One can then build \( N \) linear combinations of assets such that their returns \( \hat{r}_a \) are all uncorrelated and of unit variance. This choice is not, however, unique: in fact, any further rotation in the space of assets (ie, an orthogonal combination of the synthetic asset returns \( \hat{r}_a \)) leads to another set of uncorrelated, unit variance assets (see below). Among this infinite choice of potential “factors”, is there any one that stands out which would justify applying a maximum diversification criterion among these special assets? This is the path followed in, for example, Meucci et al (2015), where the further notion of “minimum torsion bets” was introduced.

### 3 SYMMETRIES

We want to propose a related but different route based on symmetry arguments, which fully exploits rotation and dilation invariance at the level of indicators as well as the level of returns. First, let us note that one can rescale the returns of each asset \( i \) by an arbitrary factor without changing the portfolio allocation problem. Investing 1 in a stock is the same as investing \( \frac{1}{2} \) on a fictitious “two-stock” contract, with twice the returns of the original stock. So, we can always choose to work with returns with unit variance, a choice that we will make henceforth. In this case, the covariance matrix \( C \) is, in fact, the correlation matrix between stocks. Now, the linear transformation

\[
\hat{r}_i = \sum_j (C^{-1/2})_{ij} r_j
\]

is such that \( \mathbb{E}[\hat{r}_i \hat{r}_j] = \delta_{ij} \), ie, to a set of uncorrelated assets. Here, \( C^{-1/2} \) is defined as the positive-definite square root of \( C \), namely

\[
C^{-1/2} = \sum_a \frac{1}{\sqrt{\lambda_a}} v_a v_a^T,
\]

where \( \lambda_a \) and \( v_a \) are the eigenvalues and eigenvectors of \( C \). Throughout the paper, this is the meaning we will give to the square root of a symmetric matrix. As noted above, there is a large degeneracy in the construction of the set of uncorrelated assets: any rotation of \( \hat{r} \) would do. A natural choice at this point is to insist that the \( \hat{r}_i \) are

---

1 Here and below, we assume that any nonzero average return (coming, for example, from predictive signals) is small compared with the volatility, and can be neglected in our discussion. Of course, this nonzero average return is still what motivates the portfolio construction to begin with.

2 What we call “rotations” in this paper includes proper and improper rotations, ie, rotations plus inversions.
as close as possible to the original normalized returns, so that the financial intuition about the resulting synthetic assets is preserved (to wit, $\hat{r}_i \approx \hat{r}$), This is the case for the $\hat{r}_i$ defined in (3.1). This can be shown by writing $\hat{r} = R C^{-1/2} r$, where $R$ is a rotation matrix, and demanding that any $C$ dependent on the distance between $r$ and the uncorrelated $\hat{r}$ is minimized. This leads to $R = I$.

The same construction can be applied for statistical indicators of future returns that we call $p_i$, $i = 1, \ldots, N$. We insist that $p_i$ is not necessarily the “true” expectation value of the future $r_i$, but simply the best guess of the investor based on their information, skill set, biases, etc. The standard example considered below is a trend indicator based on a moving average of past returns, but any quantitative indicator based on information or intuition would do. These indicators fluctuate in time and are also characterized by some covariance matrix $Q_{ij} = \mathbb{E}[p_i p_j]$. This matrix is, in general, nontrivial, as one may systematically predict similar returns for two different assets $i$ and $j$, leading to $Q_{ij} > 0$. In any case, one can, as above, build $N$ uncorrelated linear combinations of indicators, given by

$$\hat{p}_i = \sum_j (Q^{-1/2})_{ij} p_j,$$

with the above interpretation for $Q^{-1/2}$. The $\hat{p}_i$ are then all uncorrelated and of unit variance, i.e., with the same scale of predictability in all directions, and as close as possible to the original $p_i$, which is, again, financially meaningful. At this stage, any rotation in the space of (synthetic) assets also rotates the new indicators $\hat{p}_i$, while keeping them all uncorrelated and of unit variance. The portfolio construction problem has thus become completely isotropic.

4 ROTATIONALLY INVARIANT PORTFOLIOS

How does all this help us to construct a truly agnostic risk parity portfolio, with no reference to a specific set of assets deemed fundamental? A simple observation is that the realized gain $\hat{g}_\alpha$ of a portfolio invested in the synthetic asset $\alpha$ proportionally to $\hat{p}_\alpha$ is given by

$$\hat{g}_\alpha = \sum_{\alpha=1}^{N} \hat{p}_\alpha \cdot \hat{r}_\alpha := \sum_{\alpha=1}^{N} \hat{g}_\alpha.$$

3 The choice of normalization for the returns $r$ is important here. Indeed, working with nonnormalized returns would lead to a different result for $\hat{r}$. The choice we made is in line with our isotropy assumption.

4 The usual case of static “long-only” indicators is special, since the corresponding correlation matrix is ill-defined. This will be the subject of a forthcoming work.

5 This implicitly assumes that the cross-correlations between the $\hat{p}_\alpha$ and the $\hat{r}_\beta \neq \alpha$ are small, which is an important hypothesis underlying our rotational symmetry principle.
This portfolio has several very desirable properties.

- The risk associated with each synthetic asset is the same:

\[ \mathbb{E}[\hat{g}_\alpha^2] = \mathbb{E}[\hat{\rho}_\alpha^2] \mathbb{E}[\hat{r}_\alpha^2] = 1, \]

provided one neglects \( \mathbb{E}[\hat{g}_\alpha] \) (see footnote 1).

- The gains associated with different synthetic assets are uncorrelated:

\[ \mathbb{E}[\hat{g}_\alpha \hat{g}_\beta] = \delta_{\alpha,\beta} \]

(see footnote 5).

- Most importantly, the total gain \( \hat{g} \) is invariant under any further simultaneous rotation \( R \) of the assets and the indicators, as it should be for a scalar product:

\[
\hat{g}_{R} = \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} \mathbb{R}_{\alpha,\beta} \hat{p}_\beta \cdot \sum_{\gamma=1}^{N} \mathbb{R}_{\alpha,\gamma} \hat{r}_\gamma \\
= \sum_{\beta=1}^{N} \sum_{\gamma=1}^{N} \hat{p}_\beta \cdot \hat{r}_\gamma \sum_{\alpha=1}^{N} \mathbb{R}_{\alpha,\beta} \mathbb{R}_{\alpha,\gamma} \\
= \sum_{\beta=1}^{N} \hat{p}_\beta \cdot \hat{r}_\beta \equiv \hat{g}, \tag{4.2}
\]

where we have used the fundamental property of rotation matrixes \( RR^T = I \).

The last property means that any arbitrary choice of uncorrelated, unit variance synthetic assets with its corresponding set of indicators leads to the very same gain, so one does not need to decide on supposedly more fundamental investment factors.

Why should one invest in the synthetic asset \( \alpha \) proportionally to \( \hat{p}_\alpha \)? On the basis of symmetry arguments, this is the only rational choice. All investment directions are made statistically equivalent; any other choice would correspond to an arbitrary breaking of isotropy. In the language of Markowitz optimization, this corresponds to the optimal portfolio of synthetic assets when the expected future return of \( \alpha \) is \( S \hat{p}_\alpha \), where the expected Sharpe ratio \( S \) is independent of \( \alpha \). Note that this relies on the assumption that \( \mathbb{E}[\hat{p}_\alpha \hat{r}_\beta] = S \delta_{\alpha,\beta} \), ie, that at the level of uncorrelated factors, there is no significant cross-prediction left. This is, we believe, a very plausible assumption in practice (see below).
Now, we need to convert the above isotropic risk portfolio invested in synthetic assets into tradeable contracts. This simply follows from the definition of $\tilde{r}_a$ and $\tilde{p}_a$:

$$
\tilde{r}_a = \sum_{i=1}^{N} \sum_{j=1}^{N} (Q^{-1/2})_{aj} p_j (C^{-1/2})_{ai} r_i
$$

$$
= \sum_{i=1}^{N} \left( \sum_{a=1}^{N} \sum_{j=1}^{N} (C^{-1/2})_{ai} (Q^{-1/2})_{aj} p_j \right) r_i
$$

$$
:= \sum_{i=1}^{N} \pi_i r_i, \tag{4.3}
$$

where the last equation defines the physical position $\pi_i$ in an asset $i$, which is thus found to be

$$
\pi_i = \omega \sum_{a=1}^{N} \sum_{j=1}^{N} (C^{-1/2})_{ai} (Q^{-1/2})_{aj} p_j, \tag{4.4}
$$

where $\omega$ is a constant that sets the overall risk of the portfolio, or, in vectorial form (using the symmetry of $C$),

$$
\pi = \omega C^{-1/2} Q^{-1/2} \mathbf{p} \tag{4.5}
$$

This is the central result of this paper, which we shall now comment on and specialize to several situations. First, note that the above portfolio construction is such that the expected risk along any eigen-direction of $C$ is the same: hence the name “eigenrisk parity portfolio” (ERP). On this topic, see also Partovi and Caputo (2004) and Kind (2012). Indeed, the expected risk along the $a$th principal component is given by

$$
R_a = \mathbb{E}[(\pi \cdot v_a)^2] \lambda_a, \tag{4.6}
$$

where $\lambda_a$ is the $a$th eigenvalue of $C$, and $v_a$ is the corresponding eigenvector. Simple algebra then leads to

$$
R_a = \omega^2 \left( \frac{1}{\sqrt{\lambda_a}} \right)^2 \mathbb{E}[(v_a \cdot Q^{-1/2} \mathbf{p})^2] \lambda_a = \omega^2 \quad \text{for all } a, \tag{4.7}
$$

where we have used the fact that the expected covariance of the indicator is $Q$. Note that although for any given day the allocation $\pi$ points in a specific direction and is thus “fully concentrated” in that sense, this direction is expected to change over time, provided the indicators themselves are not static. Isotropy is thus statistically restored on long enough time scales.
5 AGNOSTIC RISK PARITY

The naive choice for the indicator covariance matrix $Q$ should be proportional to the return covariance matrix itself, ie, $Q \propto C$. In a stationary world, where the indicators would really statistically predict future returns, ie, $p_i = \mathbb{E}[r_i^{\text{fut}}]$, this assumption would be natural, at least when $C$ is computed on the time scale of the predicted returns, which is usually much longer than a day. Interestingly, plugging $Q \propto C$ into (4.5) above leads to the standard Markowitz optimal portfolio: $\pi = \omega C^{-1} \mathbb{E}[r^{\text{fut}}]$. However, this is a highly over-optimistic view of the world that only deals with “known unknowns”. Directional predictions are extremely uncertain, much more so than risk predictions. In fact, directional predictions should not even be possible in an efficient market. If one insists that some signals may (weakly) predict future returns, it is wiser not to assume any particular structure on the correlation matrix of these indicators that any optimizer would use to hedge some bets with other bets. The most agnostic choice, less prone to unknown unknowns, is $Q := \sigma_p \mathbb{I}$, ie, no reliable correlations between the realized predictions, and the same amount of predictability (or expected Sharpe ratio) on all assets. If we choose the best (RIE) estimator of the covariance matrix $C = C_{\text{RIE}}$, this leads us to a very interesting portfolio construction, namely,

$$\pi^* = \omega C_{\text{RIE}}^{-1/2} p$$

(5.1)

This allocation will be coined “agnostic risk parity” (ARP) because it allows one to precisely balance the risk between all the principal components of the (cleaned) covariance matrix $C_{\text{RIE}}$ in the worst-case scenario, where the realized correlations between indicators would completely break down.

Note that there is no explicit optimization used in this argument; rather, we look for a rotationally invariant portfolio construction with the minimal amount of information on the correlation structure of the indicators. The risk distribution per eigenmode for various portfolio allocations is shown in Figure 1, where the realized covariance of the indicators is $Q = \sigma_p \mathbb{I}$. Note that, as is well known, the Markowitz optimization scheme tends to over-allocate on small eigenmodes. This can lead to significant out-of-sample (bad) surprises (Michaud 1998), a bias that is corrected within the ARP framework.

Finally, one might believe that, although uncertain, part of the return correlations could be inherited by the indicators. A simple way to encode this is to use a shrinkage estimator for $Q$, ie, $Q \propto \varphi C_{\text{RIE}} + (1 - \varphi) \mathbb{I}$, where $\varphi \in [0, 1]$ allows one to smoothly interpolate between complete uncertainty ($\varphi = 0$), corresponding to ARP, and the standard Markowitz prescription ($\varphi = 1$).
FIGURE 1 Realized risk carried by different eigenmodes resulting from three portfolio constructions: $1/N$ on futures contracts, Markowitz and ARP, all in the case where indicators are such that their realized covariance is $Q = \sigma_p$.1.

6 AGNOSTIC TREND FOLLOWING

The previous discussion was rather formal. Here, as an example, we consider the universal “trend” indicator, based on a one-year flat moving average of past returns of a collection of 110 futures contracts (commodities, foreign exchange, indexes, bonds and interest rates; see the discussion in Lempérière et al (2014)). We normalize the returns of all futures and all the predictors to have unit variance. We then use four different portfolio constructions: an equal $1/N$ risk on each physical asset; a Markowitz optimal portfolio with either the raw empirical correlation matrix or a cleaned version $C_{RIE}$ (using the RIE estimator detailed in Bun et al (2017), and no future information); and the ARP, again using the RIE estimator for $C_{RIE}$. The profits and losses (P&Ls) of the different portfolios since 1998 are shown in Figure 2. While part of the improvement comes – as expected – from using a cleaned correlation matrix, we see that ARP yields the best result. The true correlation of predicted yearly returns $Q$ is nearly impossible to measure without centuries of data, hence motivating the choice $Q = \sigma_p$.1. We have observed similar results for other standard classic technical (CTA) strategies.
FIGURE 2  P&L curves for the universal trend following of four portfolio constructions: $1/N$ on futures contracts, Markowitz with or without a cleaned RIE correlation matrix and ARP, again with RIE.

The universe here is composed of 110 contracts (commodities, foreign exchange, indexes, bonds and interest rates). The trend indicator is a one-year flat moving average of past returns. All P&Ls are rescaled such that their realized volatility is the same.

7 PERSPECTIVES

In summary, we have offered a new perspective on portfolio allocation, which avoids any explicit optimization and instead takes the point of view of symmetry. In a context where linear combinations of assets can easily be synthesized in a portfolio whose risk is measured through volatility, the asset space can be made fully “isotropic”, in the sense that no preferred directions (corresponding to specific risk factors) can be identified. Therefore, in the absence of extra information, portfolio construction should respect this symmetry. Implementing this sole requirement leads to a precise allocation formula, (4.5), which generalizes Markowitz prescription so as to take into account the expected correlation between the predicted returns of each asset in the portfolio. We have argued that the most agnostic, and probably the most robust out-of-sample, choice is to assume that these correlations are zero, i.e., one should refrain from trying to hedge different bets if there is no certainty about the correlations between these bets. This leads to an ARP portfolio that realizes an equal risk over all principal
components of the covariance matrix. We found that such an allocation outperforms Markowitz’s portfolios when applied to CTA strategies, such as (universal) trend following. There are several routes that should be explored further. For example, nonquadratic measures of risk, such as skewness or kurtosis, would break rotational symmetry and possibly lead to meaningful fundamental risk factors that could be maximally diversified (see, for example, Baitinger et al 2017). We leave this for future work.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

We thank N. Bercot, J. Bun, R. Chicheportiche, S. Ciliberti, C. Deremble, L. Duchayne, L. Laloux, A. Rej and J. Zuniga for many useful discussions on these issues.

REFERENCES


Research Paper

Black–Litterman, exotic beta and varying efficient portfolios: an integrated approach

Ricky Alyn Cooper¹ and Marat Molyboga²

¹Illinois Institute of Technology, 565 West Adams Avenue, Chicago, IL 60661, USA; email: rcooper3@iit.edu
²Efficient Capital Management, 4355 Weaver Parkway, Warrenville, IL 60555, USA; email: molyboga@efficient.com

(Received October 24, 2016; accepted January 16, 2017)

ABSTRACT

This paper brings Black–Litterman optimization, exotic betas and varying starting portfolios together into one complete, symbiotic framework. The approach is unique because these techniques are often viewed as alternatives rather than as complements to each other. We first demonstrate the approach using exotic beta as the “views” in the Black–Litterman optimization. This framework benefits investors who already utilize the classic Black–Litterman approach and appreciate advances in the exotic beta research, and also those who focus on practical implementation of exotic betas. We then explore the framework using the risk-parity portfolio as an efficient starting portfolio for Black–Litterman optimization on both theoretical and practical grounds. We demonstrate that risk parity is a highly effective starting point in many situations. Finally, as part of our discussion, we derive conditions under which almost any completely diversified portfolio may be used as a starting portfolio in the Black–Litterman process. The integrated methodology developed is robust, flexible and easily implemented, which means that a wide range of investors can benefit from this framework.

Keywords: Black–Litterman; exotic betas; risk parity; reverse optimization; alpha model; risk-adjusted returns.
1 INTRODUCTION

The Black–Litterman model has contributed to the field of quantitative portfolio management by elegantly applying Bayesian statistics to marry two seemingly contradictory ideas: the efficiency of the market portfolio and the efficacy of alpha modeling. However, the practical implementation of the model is often difficult because it relies on expert opinions, which often are hard to obtain and of uncertain quality, and the capitalization weights of the market portfolio, which are not always available. Further, the market portfolio is assumed to be nearly efficient. The latter assumption has often been questioned over the last two decades.

Exotic beta is a well-established concept emanating from a large body of research that presents a ready-made set of prior beliefs for people who trust the established literature more than a privately built model. This paper explores two aspects of Black–Litterman and exotic beta. First, we consider the likelihood that exotic beta will improve the performance of an index portfolio inside a Black–Litterman framework. Second, we consider using the Black–Litterman framework to derive an implementation portfolio for exotic beta. Our conclusion is that this combination is effective on both counts.

From a practical point of view we wish to address situations that may come up in the real world.

(1) A manager likes the Black–Litterman framework but does not have access to a good proprietary view. In this case, exotic beta may perform as the options.

(2) A manager likes the idea of the Black–Litterman framework but does not necessarily believe that a capitalization-weighted portfolio is efficient, or cannot get accurate capitalization weights. In this case, the risk-parity portfolio may perform as a starting portfolio.

(3) A manager likes exotic beta tilts to a portfolio and seeks a good implementation methodology. In this case, the Bayesian construction of Black–Litterman scales the exotic beta and provides returns that can be optimized.

(4) A portfolio manager wishes to use the Black–Litterman model, but does not wish, or is unable, to start with the market portfolio. In this case, we derive a more general reverse optimization that can be used with any well-diversified portfolio.

The net result of this research is that a wide range of investment professionals – including the portfolio manager interested in applying exotic betas, the fund-of-funds manager or commodity trading advisor interested in applying Black–Litterman and anyone interested in extending their tool set of allocation techniques – should find these results appealing.
The paper unfolds as follows. First, we demonstrate the manner in which exotic betas may be integrated with the Black–Litterman framework using a simple ten-stock example with an exotic beta of low volatility. Second, we illustrate the difference between starting Black–Litterman with a risk-parity portfolio and starting it with a capitalization-based portfolio using the same ten-stock example. Finally, we demonstrate how Black–Litterman, risk parity and exotic beta can be integrated within the Bayesian risk-parity framework using a three-asset-class example of stocks, bonds and commodities, which is particularly interesting because capitalization weights are not available. For this example, cross-sectional momentum is chosen as the exotic beta.

Along the way, we will also demonstrate an algorithm by which almost any very well-diversified portfolio may be used as the starting portfolio in the Black–Litterman framework, without sacrificing theoretical integrity.

2 STANDARD BLACK–LITTERMAN OPTIMIZATION

The original Black–Litterman model changed the landscape of quantitative portfolio management by combining into a single framework the two seemingly contradictory ideas of the efficiency of the market portfolio and alpha models. Black–Litterman optimization takes the implied returns from a cap-weighted index, which represents the market portfolio, combines them with a personal view on expected returns (the prior) and reinverts the linear combination of expected returns (the posterior) into a final portfolio.

The capital asset pricing model (CAPM; see Sharpe 1964; Lintner 1965) suggests that the market portfolio is an excellent starting point because, under fairly restrictive assumptions, in equilibrium the expected return from diversifiable risk is equal to zero and the market portfolio should have the highest Sharpe ratio. However, Black–Litterman also allows personal views on expected returns for assets that are deemed to be away from their equilibrium value. As we will show, the Black–Litterman solution represents returns that are a weighted average of the market portfolio (or, more generally, the data model portfolio) and a portfolio that

---

1 Roll (1978) dimmed some hopes by pointing out that the ultimate market portfolio is unobservable in the real world, but he also pointed out that any very well-diversified cap-weighted portfolio would have little unsystematic risk, and that portfolios of these indexes would be nearly efficient and still have returns directly related to their betas. These insights led to the philosophy that, with the preponderance of people all keeping the market efficient through analysis, trying to beat the market was a fool’s errand. As Ellis (1975) stated, indexing and long-term goal planning were the way to avoid playing a “loser’s game”.

www.risk.net/journals

To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
fully incorporates the private expected return model (the prior). This suggests that the many desirable properties of the Black–Litterman methodology, documented in Black and Litterman (1992) and Bevan and Winkelmann (1998), are driven by diversification.

As Markowitz (1952) suggests, the efficient frontier with a risk-free asset available is defined by a linear function of portfolio weights, $w^*$, that solves the problem

$$
\min_{w} \frac{1}{2} w'Vw - \lambda (T - r_f - (r - r_f)'w),
$$

(2.1)

where $r$ is a vector of expected returns, $w$ is the vector of portfolio weights, $V$ is the covariance matrix of excess returns, $\lambda$ is a Lagrange multiplier closely related to the inverse of risk aversion, $T$ is a target return and $r_f$ is the risk-free rate. It is easily shown that the capitalization-weighted portfolio is a solution to (2.1), assuming the assumptions of the CAPM are not violated. This vector of returns (the starting solution) may be expressed as

$$
r^* = r_f e + \left( \frac{r_m - r_f}{\sigma_{r_m}} \right)^{-1} \sigma_{r_m} V w^*,
$$

(2.2)

where $r_m$ is the capitalization-weighted (market) return, $e$ is a conformable vector of ones and $w^*$ is the vector of capitalization weights. Note that the term in parentheses represents the Sharpe ratio. It may be estimated from the historical data, or from other outside considerations. Its appearance comes from the fact that, for a target rate in (2.1) equal to the market return,

$$
\lambda = \left( \frac{r_m - r_f}{\sigma_{r_m}} \right)^{-1}
$$

is a necessary part of the solution.

Equation (2.2) gives returns consistent with the capitalization-weighted portfolio being on the minimum variance set, along the tangency line from the risk-free rate, and is used in deriving implied returns of the Black–Litterman model. Following Black and Litterman (1992) and Meucci (2010), we allow the possibility that, besides the “public” opinion of returns (public in that it is based on the market being in equilibrium), a portfolio manager gives credence to a private model of returns: a prior belief that some additional factors are needed to determine the equilibrium return. The Black–Litterman model imposes such beliefs using a matrix of constraints on the distribution of returns. For example, if we wanted to model the belief that all returns would be equal next period (we are not advocating this), it could be written within
the standard Black–Litterman framework as

\[ P_{r^{**}} = \begin{bmatrix}
1 & -1 & 0 & \cdots & 0 & 0 \\
0 & 1 & -1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -1 & 0 \\
0 & 0 & 0 & \cdots & 1 & -1 \\
\end{bmatrix} \begin{bmatrix}
r_{1}^{**} \\
r_{2}^{**} \\
\vdots \\
r_{n-1}^{**} \\
r_{n}^{**} \\
\end{bmatrix} \sim N(\nu, \Omega). \tag{2.3}
\]

Equation (2.3) implies that the difference in the forecasted returns of any two assets is distributed normally around some values, which in this case would be a vector of zeros. This particular set of beliefs is that the difference in return of any two assets will be expected to be zero with some variability.

To account for the variability, a good choice is to follow Meucci (2010) and set

\[ \Omega = \frac{1}{c} P V P^T, \]

where \( c \) represents the strength of conviction about the prior.\(^2\)

The brilliance of Black–Litterman optimization is that the final (posterior) return vector may be written as

\[ r = (1 - x)r^* + x r^{**}, \tag{2.4} \]

where

\[ x = \frac{c}{1 + c} \quad \text{and} \quad r^{**} = r^* + V P^T (P V P^T)^{-1} (\nu - P r^*). \]

Equation (2.4) implies that the posterior return is a portfolio of the data-model-based return and the prior-based return. This also means that the portfolio formed from the posterior returns will be a linear combination of the portfolios formed from the data-based returns and prior returns. Moreover, the portfolio resulting from the posterior returns (by (2.2)) will also be a linear combination of the portfolios obtained from the starting returns and the views-based returns.\(^3\) Since these portfolios will not generally have perfect correlation, this yields diversification benefits that can potentially

---

\(^2\) A \( c \) near zero would mean no belief in the prior (and thus it should not be considered), whereas \( c \) tending toward infinity would indicate certainty that the prior is the correct forecast for the next period (this return is a conditional return, based on the state of the efficient frontier, even if we think it is, at least in part, completely incorrect).

\(^3\) One minor problem is that the Bayesian return vector might not reinvert into a portfolio with normalized weights using (2.2) if assets are highly correlated. In this paper, we simply rescale the weights to equal 1 in total. An alternative would be to take the returns and apply them to Markowitz optimization or a more general quadratic program directly.
improve the risk-adjusted return of the data portfolio, as long as the prior returns are sufficiently strong.  

3 THE BLACK–LITTERMAN FRAMEWORK WITH EXOTIC BETAS

What became known as “exotic beta” appeared in the literature in the very early days of equilibrium asset pricing theory. Deriving from the literature originally referred to as the anomaly literature, meaning things that are outside the standard canon of CAPM orthodoxy, such non-CAPM risk factors began to really take shape in work by Fama and French (1992) and, later, Carhart et al (1997). Carhart et al (2014) explored the notion of exotic beta rather fully and concluded that exotic beta is a powerful portfolio management tool. They defined exotic betas as exposures to risk factors that are uncorrelated with global equity markets and have positive expected returns. This definition suggests that exotic betas are not stock-specific “alphas” in the traditional sense of being an unsystematic return specific to a stock, but rather are sensitivities to risk premiums that are outside of the CAPM, and presumably more stable than the market risk premium.

Exotic betas (or any other “alpha” model) might seem to invalidate the Black–Litterman framework, which assumes that the market portfolio is nearly efficient. If the CAPM were strictly true, the market portfolio would account for all systematic risk, and no prior return vector (due to exotic beta, or otherwise) would be needed. So, when we are using a Black–Litterman framework, the starting index portfolio may be thought of as approximately on the efficient frontier, but there is some inefficiency (or weighted group of inefficiencies) that can be added to the market returns to make more accurate asset return forecasts.

Another interpretation of the Black–Litterman approach is to treat it as a mixture model, where we believe there is a \(1/(1+c)\) chance that the market is efficient by itself and the implied expectations are correct, and a \(c/(1+c)\) chance that a different set of return expectations is the efficient portfolio. The combination of the two probabilities results in hybrid forecasts of asset returns that are part market portfolio return and part exotic beta return. Of course, many methods of applying exotic betas are possible, but the additional benefit of this method is that it provides a distinctly different approach to extracting returns from exotic betas from those documented in the literature.

The nice thing about (2.3) is that it allows the reverse-optimization returns to scale the exotic beta returns. Looking closely at (2.3), we note that there are \(n-1\) equations and \(n\) assets. This allows infinitely many solutions. However, (2.4) specifies that the

\[ \text{Sharpe}(y) > \text{corr}(x, y) \times \text{Sharpe}(x). \]

Though the CAPM suggests that there should be no priors strong enough to improve performance, a rich literature on market anomalies presents evidence that such priors exist.

---

4 The required strength of the prior returns can be expressed using a simple inequality. Prior returns will improve performance of the data-based returns if and only if \(\text{Sharpe}(y) > \text{corr}(x, y) \times \text{Sharpe}(x)\). Though the CAPM suggests that there should be no priors strong enough to improve performance, a rich literature on market anomalies presents evidence that such priors exist.
solution chosen will be the one that puts the exotic beta returns on the same scale as the reverse-optimization returns.

To make this discussion more concrete, Table 1 gives descriptive statistics for ten Dow Jones industrial stocks, whose returns are observed monthly, for the period from January 1995 to May 2015. This small number of stocks was chosen to make covariance estimation a trivial issue that would not interfere with the main points of the paper.\(^5\)

We illustrate how to incorporate exotic betas in the Black–Litterman framework by starting with a capitalization-weighted portfolio of the ten stocks and using the low-volatility anomaly, described in Jagannathan and Ma (2003), as an example of an exotic beta. In this case, the future expected Sharpe ratio is considered to be an inverse function of previous volatility. An easy way to represent this prior within the current framework is to state, for any pair of assets, the expected difference in their Sharpe ratios going forward as a percentage of the trailing difference in their inverse volatility.

Specifically, this prior return, consistent with the low-volatility anomaly, may be written for all assets as

\[
P r^{**} = \begin{bmatrix}
\frac{1}{\sigma_1} & -\frac{1}{\sigma_2} & 0 & \cdots & 0 & 0 \\
0 & \frac{1}{\sigma_2} & -\frac{1}{\sigma_3} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -\frac{1}{\sigma_{n-1}} & 0 \\
0 & 0 & 0 & \cdots & \frac{1}{\sigma_{n-1}} & -\frac{1}{\sigma_n}
\end{bmatrix}
\times
\begin{bmatrix}
r_1^{**} \\
r_2^{**} \\
\vdots \\
r_{n-1}^{**} \\
r_n^{**}
\end{bmatrix}
\sim N(\nu, \Omega),
\tag{3.1}
\]

with

\[
v_j = \alpha \left( \frac{1}{\sigma_i} - \frac{1}{\sigma_{i+1}} \right).
\]

For this example we use alpha equal to 0.001, and the prior return equation indicates there is an expectation that the difference in Sharpe ratio between two assets will be directly related to the difference in their inverse volatilities.

---

\(^5\) There is no consensus as to the optimal estimation method for larger covariance matrixes, but a number of approaches have been introduced. Ledoit and Wolf (2004) suggest shrinking a sample covariance matrix; Jagannathan and Ma (2003) consider using daily returns and factor models in addition to shrinkage; Pafka et al (2004) argue for applying filtering based on the random matrix theory. The issue of estimating the covariance matrix is beyond the scope of this paper. For this particular small problem, covariance estimates rely on a simple sixty-month sample.
TABLE 1  Descriptive statistics of the ten Dow Jones 30 stocks.

<table>
<thead>
<tr>
<th>Stocks</th>
<th>Annual excess return (%)</th>
<th>Annual SD (%)</th>
<th>Sharpe ratio</th>
<th>HD</th>
<th>DIS</th>
<th>NKE</th>
<th>MCD</th>
<th>PG</th>
<th>WMT</th>
<th>KO</th>
<th>CVX</th>
<th>XOM</th>
<th>TRV</th>
</tr>
</thead>
<tbody>
<tr>
<td>HD</td>
<td>10.63</td>
<td>26.94</td>
<td>0.39</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIS</td>
<td>7.83</td>
<td>25.72</td>
<td>0.30</td>
<td>0.33</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NKE</td>
<td>13.74</td>
<td>30.31</td>
<td>0.45</td>
<td>0.32</td>
<td>0.28</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCD</td>
<td>8.15</td>
<td>21.34</td>
<td>0.38</td>
<td>0.31</td>
<td>0.37</td>
<td>0.34</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PG</td>
<td>7.48</td>
<td>20.29</td>
<td>0.37</td>
<td>0.09</td>
<td>0.10</td>
<td>0.16</td>
<td>0.25</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMT</td>
<td>8.06</td>
<td>22.13</td>
<td>0.36</td>
<td>0.51</td>
<td>0.13</td>
<td>0.35</td>
<td>0.30</td>
<td>0.16</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO</td>
<td>5.19</td>
<td>21.40</td>
<td>0.24</td>
<td>0.14</td>
<td>0.30</td>
<td>0.26</td>
<td>0.40</td>
<td>0.46</td>
<td>0.26</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVX</td>
<td>8.32</td>
<td>19.87</td>
<td>0.42</td>
<td>0.25</td>
<td>0.32</td>
<td>0.25</td>
<td>0.37</td>
<td>0.07</td>
<td>0.15</td>
<td>0.25</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>XOM</td>
<td>8.34</td>
<td>16.74</td>
<td>0.50</td>
<td>0.14</td>
<td>0.29</td>
<td>0.16</td>
<td>0.34</td>
<td>0.22</td>
<td>0.08</td>
<td>0.33</td>
<td>0.72</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>TRV</td>
<td>7.20</td>
<td>25.58</td>
<td>0.28</td>
<td>0.31</td>
<td>0.36</td>
<td>0.37</td>
<td>0.32</td>
<td>0.13</td>
<td>0.26</td>
<td>0.34</td>
<td>0.37</td>
<td>0.30</td>
<td>1.00</td>
</tr>
</tbody>
</table>

TABLE 2  Black–Litterman applied to the ten-stock example using the low-volatility anomaly.

<table>
<thead>
<tr>
<th></th>
<th>$c = 0$</th>
<th>$c = 1/3$</th>
<th>$c = 1$</th>
<th>$c = 3$</th>
<th>$c = \infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized excess return (%)</td>
<td>3.64</td>
<td>3.92</td>
<td>4.21</td>
<td>4.50</td>
<td>4.79</td>
</tr>
<tr>
<td>Annualized standard deviation (%)</td>
<td>12.17</td>
<td>12.15</td>
<td>12.45</td>
<td>13.10</td>
<td>14.08</td>
</tr>
<tr>
<td>Sharpe ratio</td>
<td>0.299</td>
<td>0.323</td>
<td>0.338</td>
<td>0.344</td>
<td>0.340</td>
</tr>
</tbody>
</table>


TABLE 3  Comparison of two implementations of the low-volatility anomaly.

<table>
<thead>
<tr>
<th></th>
<th>Standard implementation</th>
<th>Bayesian prior approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized excess return (%)</td>
<td>0.68</td>
<td>1.15</td>
</tr>
<tr>
<td>Annualized standard deviation (%)</td>
<td>11.78</td>
<td>8.39</td>
</tr>
<tr>
<td>Sharpe ratio</td>
<td>0.06</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Standard refers to a typical long–short implementation. The Bayesian prior approach is a long–short portfolio that goes long Black–Litterman with infinite $c$ (100% weight to the low-volatility prior) and short market portfolio. Out-of-sample period: February 2000 to May 2015.

Table 2 outlines the results of applying the Black–Litterman model to a market-weighted reverse-optimization model and a low-volatility prior with several levels of $c$ that represent 25%, 50%, 75% and 100% weights to the low-volatility anomaly prior. In this case, the maximal Sharpe ratio of 0.344 is accomplished at $c = 3$, representing a 75% weight to the prior, which is superior to 0.299, the Sharpe ratio of the market portfolio; this suggests that the low-volatility anomaly view used within the Black–Litterman framework can substantially improve performance.

Another interesting experimental result is shown in Table 3, which compares the performance of the standard implementation of the low-volatility anomaly and our implementation, based on the Black–Litterman framework. The standard implementation ranks stocks based on their in-sample volatility and then buys the bottom quintile of stocks (low-volatility stocks) and sells short the top quintile of stocks (high-volatility stocks). Portfolio weights are inversely proportional to historical inverse volatilities. The Black–Litterman approach purchases the highest Sharpe ratio portfolio implied by Black–Litterman returns with $c = \infty$ (100% weight to the low-volatility anomaly prior) and sells short the market portfolio.

6 For example, if $c = 1/3$, this represents a $1/(1+c) = 25\%$ weight assigned to return expectations of the market portfolio and a 75% weight assigned to those of the prior.

7 Once we have the Black–Litterman returns we can use standard Markowitz optimization or constrained quadratic optimization. The results in Table 3 use standard Markowitz optimization to find the highest forecast Sharpe ratio portfolio.
The Sharpe ratio of the Bayesian prior approach is equal to 0.14, which is more than twice 0.06, the Sharpe ratio of the standard implementation of the low-volatility anomaly. Though the relative performance of the two implementations of the low-volatility anomaly (or any exotic beta in general) can be sensitive to the time period, portfolio constituents and choice of parameters, the Bayesian prior approach substantially expands the toolbox of exotic beta implementation with potentially significant performance implications for investors.

Carhart et al (2014) suggest that utilizing a limited version of Black–Litterman with exotic betas as portfolio constituents is unlikely to diminish its power. We have extended this result by showing that the standard Black–Litterman implementation can be combined with exotic betas, by using them as views, to achieve multiple useful results. In the next sections, we extend the notion of reverse optimization to utilize portfolios other than the capitalization-weighted portfolio. This paper is, to the best of our knowledge, the first to investigate the implications of alternative efficient portfolios for Black–Litterman optimization and suggest a version of Black–Litterman optimization that extends its application to many new investment situations.

4 THE MARKET PORTFOLIO, THE RISK-PARITY PORTFOLIO AND EFFICIENCY

While the CAPM suggests that a capitalization-weighted market portfolio should have the highest Sharpe ratio, there are situations in which the market portfolio is either suboptimal or even inappropriate. For example, fund-of-hedge-funds allocation decisions, and decisions involving futures contracts generally, do not readily admit capitalization calculations, which makes capitalization-weighted market portfolios unattainable. Moreover, Asness et al (2012) suggest that the market portfolio is not efficient, and instead argue that the risk-parity portfolio approach, which equalizes the contribution to portfolio risk from each constituent, is more efficient due to leverage aversion.8 Qian (2006) provides a comprehensive analysis of risk-parity portfolios. We outline the technical details of the risk-parity approach in the online appendix.

Since capitalization weights are available in the ten-stock Dow Jones example, we can easily compare the performance of the capitalization-weighted market portfolio and risk parity. Table 4 reports the out-of-sample performance of the two portfolios.

The risk-parity portfolio delivers a Sharpe ratio of 0.45, which is higher than 0.30, the Sharpe ratio of the market portfolio. However, we need to be careful about drawing

---

8 Asness et al (2012) argue that leverage aversion changes the predictions of modern portfolio theory, because investors without access to leverage are unable to benefit from higher risk-adjusted returns of safer (low-beta or low-volatility) assets. Risk parity portfolios overweight safer assets relative to the market portfolio and benefit from their higher risk-adjusted returns after applying leverage.
conclusions about the efficiency of risk parity from this simple example. Anderson et al (2012) argue empirical studies that make claims about the efficiency of risk parity might be very sensitive to the time period studied and the transaction costs assumed. Our simple example is also not immune to their criticism.

To summarize, there is reason to believe that, at least sometimes, an equal-risk-weighted index portfolio may be more efficient than a capitalization-weighted portfolio. Also, there are times when capitalization weights are unavailable. In either of these situations, the risk-parity portfolio is a good candidate for the data-based starting point in the Black–Litterman framework.9

5 AN ALTERNATIVE REVERSE OPTIMIZATION

The importance of the Black–Litterman framework is that it provides a way of mixing market-data-based returns with prior views about assets that are not priced properly by the market. A key thing to realize is that reversing the first-order condition in the Markowitz model is theoretically permissible for any portfolio on the minimum variance set. The mathematics are only slightly more complicated, and the basics are outlined below.

The Markowitz model may be written in an alternative form as

$$\min \frac{1}{2} w' V w - \lambda (T - r^T w) + \gamma (1 - w'e).$$

In this formulation there are two constraints. The risky assets’ weights must sum to 1, and the target expected return must be achieved. By varying the target, we map the entire minimum variance set. This formulation leads to the following reverse-optimization returns:

$$r = \frac{V w_p - \gamma_p e}{\lambda_p},$$

(5.1)

where $w_p$ are the weights of the presumed efficient portfolio.

9 Although some people may find the notion that the equal risk portfolio is possibly efficient strange, even less likely is the fact that diversified portfolios have been shown to be efficient in some cases. DeMiguel et al (2009) maintain that, for their universe, the equal-weighted $1/N$ portfolio was more efficient than more conventional alternatives.
Now, to complete the solution of (5.1), we need to define $r_{mv}$ as the return of the minimum variance portfolio, whose weights are given by

$$w_{mv} = \frac{V^{-1}e}{e'V^{-1}e},$$

which is a function only of the elements of the covariance matrix, not expected returns. Now,

$$\begin{align*}
\lambda_p &= \frac{\sigma^2_p - \sigma_{mv,p}}{r_p - r_{mv}}, \\
y_p &= \lambda_p r_p - \sigma^2_p.
\end{align*}$$

(5.2)

The individual reverse-optimization returns may be obtained using (5.1) and (5.2). In this formulation, we have replaced the use of information about the risk-free rate and the capitalization-weighted portfolio with the use of information about the minimum variance portfolio and our assumed minimum variance portfolio.

Now that we have two reverse-optimization methodologies, which should we use? Figure 1 helps us answer this question. If we believe our portfolio is a minimum variance portfolio and is the one with the highest Sharpe ratio, then it is at point P1 and we should use (2.2). This is the market portfolio, if the CAPM is strictly true, but may be another portfolio if it is not. Only if the CAPM is true will P1 be the capitalization-weighted portfolio. If we believe that we have a minimum variance portfolio (mainly a very well-diversified portfolio) but not that it is at P1 (perhaps it is at P2), then we should use (5.1) and (5.2). Equations (5.1) and (5.2) are useful when the only thing we know is that our starting portfolio is very well diversified with respect to the investment universe. Equation (2.2) is useful when we believe our starting portfolio has not only the minimum variance but also the maximal Sharpe ratio.
### TABLE 5  Descriptive statistics of bonds, stocks and commodities.

<table>
<thead>
<tr>
<th></th>
<th>Bonds</th>
<th>Stocks</th>
<th>Commodities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized excess return (%)</td>
<td>2.79</td>
<td>3.36</td>
<td>−0.71</td>
</tr>
<tr>
<td>Annualized standard deviation (%)</td>
<td>5.45</td>
<td>14.63</td>
<td>19.26</td>
</tr>
<tr>
<td>Sharpe ratio</td>
<td>0.51</td>
<td>0.23</td>
<td>−0.04</td>
</tr>
</tbody>
</table>

This table shows the annualized excess return, standard deviation and Sharpe ratio of bonds, stocks and commodities for March 1976 through May 2015. The Barclays US Aggregate Government Index is used as a proxy for the bond market, The MSCI World Index is used as a proxy for the stock market, the S&P Goldman Sachs Commodities Index is used as a proxy for the commodities market and the three-month treasury bill secondary market rate is used as a proxy of the risk-free rate.

Having established in the previous section that the risk parity can be considered a reasonable maximal Sharpe ratio portfolio, we will take it as a starting point for Black–Litterman using (2.2). One interesting thing to note (discussed in the online appendix) is that the risk-parity portfolio simply says that each asset’s contribution to total risk should be made equal. This does not mean that each asset should have the same Sharpe ratio or the same expected return.

### 6  TWO EXAMPLES OF A RISK-PARITY STARTING PORTFOLIO

We illustrate the risk-parity and exotic beta approach by considering two simple examples with investments in three major asset classes: stocks, bonds and commodities. The data for these experiments covers the period from March 1976 through May 2015, and consists of the MSCI World Index, as a proxy for the stock market, the Barclays US Aggregate Government Index, to represent the bond market, and the Standard & Poor’s (S&P) Goldman Sachs Commodity Index, to represent commodities. We use the three-month treasury bill secondary market rate as a proxy for the risk-free rate. Inclusion of commodities is particularly interesting because they do not have capitalization weights, and therefore the capitalization-weighted market portfolio is unattainable.\(^{10}\)

Table 5 presents some relevant descriptive statistics of the data. One interesting statistic is that the three asset classes perform very differently over this time period, with commodities performing worst, with a Sharpe ratio of −0.04, and bonds performing best, with a Sharpe ratio of 0.51.

\(^{10}\) Portfolio allocations that involve hedge funds are another example of limitation in the capitalization-weighted approach.
However, the relative performance of the three asset classes across time is very inconsistent. Figure 2 displays the rolling twenty-four-month Sharpe ratio of the three assets.\textsuperscript{11}

Absolute and relative performance is inconsistent across time, with the Sharpe ratios ranging between around $-2.5$ and around $+2.5$.

The first step in this extended Black–Litterman approach is calculating risk-parity weights and the corresponding implied expected excess returns using (2.2). The second step involves imposing exotic-beta-based views. We consider two examples that can have broad applications to many investors. The first example considers momentum as an exotic beta that is robust across most asset classes, as documented in Asness \textit{et al} (2013). The second example considers an equal Sharpe ratio prior.

\subsection{6.1 Risk parity with momentum}

Momentum is a pervasive anomaly that has been extensively documented in the literature. We express belief in momentum in Sharpe ratios using (3.1) with the same matrix $P$ and vector

\[ v_i = \alpha \left( \frac{\tilde{R}_i}{\sigma_i} - \frac{\tilde{R}_{i+1}}{\sigma_{i+1}} \right). \]

We set $\alpha = 0.05$, which represents the belief that the difference in the future Sharpe ratios of any two assets is expected to equal 5\% of the most recent Sharpe ratios. In this simple example we use a time window of twenty-four months to estimate recent Sharpe ratios. As before, we use the same levels of $c$, representing 25\%, 50\%, 75\% and 100\% weights to the exotic beta belief. The results of this experiment are presented in Table 6.

In this case, the maximal Sharpe ratio of 0.35 occurs at $c = 3$, representing a 75\% weight to the prior, which is superior to 0.253, the Sharpe ratio of the risk-parity portfolio; this suggests that the momentum in the risk-adjusted return prior used within the Black–Litterman framework can improve performance.

\subsection{6.2 Risk parity with equal Sharpe ratios}

Although belief in equal Sharpe ratios is unrelated to exotic beta, it is an interesting case to study. There is a sizable group of investors who hold this belief about diversified asset classes. In addition, there are fund-of-(hedge) funds managers who have very high requirements for their hedge funds; this is reflected in their rigorous due diligence steps, which result in approximately the same Sharpe expectations across hedge funds.

\textsuperscript{11}In this section, we choose to use rolling twenty-four-month sample estimates for all covariance and variance estimates. With only three assets, a longer period is not required.
FIGURE 2  Rolling twenty-four-month Sharpe ratios of stocks, bonds and commodities.

<table>
<thead>
<tr>
<th></th>
<th>Stocks</th>
<th>Bonds</th>
<th>Commodities</th>
</tr>
</thead>
</table>

TABLE 6  The three-asset example with risk parity and momentum.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1/3</th>
<th>1</th>
<th>3</th>
<th>∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized excess return (%)</td>
<td>2.08</td>
<td>2.07</td>
<td>2.02</td>
<td>1.96</td>
<td>1.90</td>
</tr>
<tr>
<td>Annualized standard deviation (%)</td>
<td>8.21</td>
<td>6.52</td>
<td>5.89</td>
<td>5.59</td>
<td>5.44</td>
</tr>
<tr>
<td>Sharpe ratio</td>
<td>0.253</td>
<td>0.318</td>
<td>0.343</td>
<td>0.350</td>
<td>0.349</td>
</tr>
</tbody>
</table>

in the portfolio. Finally, it is illuminating to show how even a fairly weak prior performs in the Black–Litterman framework.

Table 5 shows that commodities substantially underperformed stocks and bonds over the time period in this study. However, other studies, such as Gorton and Rouwenhorst (2006), argue that an equally weighted index of commodity futures’ monthly returns should deliver a Sharpe ratio comparable to that of equities. The equal Sharpe belief can be expressed using (3.1) with the mean vector set to zero.

Table 7 reports results of the out-of-sample performance. The maximal Sharpe ratio of 0.316 corresponds to \( c = 1 \), representing a 50% weight to the prior, which is superior to 0.253, the Sharpe ratio of the risk-parity portfolio. This suggests that the prior of equal Sharpe ratios used within the Black–Litterman framework can add value.
TABLE 7  The three-asset example with risk parity and equal Sharpe priors.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1/3</th>
<th>1</th>
<th>3</th>
<th>∞</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annualized excess return (%)</td>
<td>2.08</td>
<td>2.03</td>
<td>1.93</td>
<td>1.81</td>
<td>1.69</td>
</tr>
<tr>
<td>Annualized standard deviation (%)</td>
<td>8.21</td>
<td>6.65</td>
<td>6.10</td>
<td>5.87</td>
<td>5.79</td>
</tr>
<tr>
<td>Sharpe ratio</td>
<td>0.253</td>
<td>0.305</td>
<td>0.316</td>
<td>0.308</td>
<td>0.292</td>
</tr>
</tbody>
</table>


FIGURE 3  Risk-parity performance

The figure displays results from Table 7. \( c = \infty \) is the point furthest south and \( c = 0 \) is the point furthest northeast, representing the classical risk-parity approach. The line from the origin is tangential to the frontier at the point with the highest Sharpe ratio, which is approximately \( c = 1 \). The axes show excess returns and excess standard deviation.

The above approach gives combinations of starting returns and view returns that allow us to reach Sharpe ratios unobtainable through either approach alone. The exact nature of this improvement is apparent in Figure 3, which presents the data in Table 7 in graphical form, and once again demonstrates the diversifying power of Black–Litterman even in the face of a rather weak prior.

The curve, reminiscent of a Markowitz frontier, reiterates the benefits of diversification produced by the Bayesian risk-parity method. The straight line from the origin to the combination line of strategic outcomes shows that the maximum obtainable Sharpe ratio is at approximately \( c = 1 \) for this particular combination of strategies and assets over this particular period. While it is true that this diversification benefit would occur with almost any prior, the better the prior actually is, of course, the better the opportunities are.
7 CONCLUSION

In this paper, we demonstrate that using the exotic beta as a prior alpha model in Black–Litterman optimization is attractive to investors that already utilize the classic Black–Litterman approach and seek to incorporate advances in the exotic beta research, and to those who focus on practical implementation of exotic betas. The reason for this behavior is that the diversification of alpha sources benefits an investor, whether they want a portfolio that is mainly efficient-portfolio based, mainly exotic-beta based or one that maximizes the Sharpe ratio.

In addition, we introduce risk parity as a valid starting portfolio, and produce a methodology for using almost any well-diversified portfolio as a starting portfolio. This is useful when the capitalization-weighted portfolio is not an appropriate starting point.

Our extended Black–Litterman approach symbiotically unifies the Black–Litterman optimization, exotic betas and risk parity into a single, flexible framework that combines the various strengths of the three approaches to improve investors’ portfolios. These results give a large number of investment professionals new tools for their investment tool box without throwing out everything they might previously have been using.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES


Research Paper

Correctness of backtest engines

Robert Löw,1 Stanislaus Maier-Paape2 and Andreas Platen2

1Department of Mathematics, KU Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium; email: robert.low@kuleuven.be
2Institut für Mathematik, RWTH Aachen, Templergraben 55, 52062 Aachen, Germany; emails: maier@instmath.rwth-aachen.de, platen@instmath.rwth-aachen.de

(Received October 25, 2016; revised February 6, 2017; accepted February 8, 2017)

ABSTRACT

In recent years, several trading platforms have appeared that provide a backtest engine to calculate the historic performance of self-designed trading strategies on underlying candle data. The construction of an accurate backtest engine is, however, a subtle task, as shown in previous work by Maier-Paape and Platen. Several platforms are struggling to achieve accuracy. We discuss how the accuracy of backtest engines can be verified and provide models for candles and intra-period prices, which may be applied to conduct a “proof of correctness” for a given backtest engine when our tests on specific model candles are successful. Further, we suggest algorithmic considerations in order to allow a fast implementation of the tests necessary for the proof of correctness.

Keywords: historical simulation; trading system; candlestick chart; imperfect data; price model; backtest correctness.

1 INTRODUCTION

There are many different market trading tools that allow programming (fully automated) trading strategies for, say, stock markets. For example, the user may be
provided with a graphical interface where a chart shows, eg, daily closing prices of a stock. It may then be possible to apply indicators and trading strategies to this chart. In this context, a trading strategy means a logic system that uses price data and decides whether or not to push signals such as sending a limit buy entry order or a stop market order to buy/sell shares of stock.

A trading strategy can be applied to historical data to find out how well the strategy would have worked in the past. Therefore, the trading software needs to simulate the trading strategy using known price data. In most cases, the software uses the displayed data, eg, the daily closing prices, and performs the historical simulation. The result of this simulation is a list of trades including, among other things, their entry and exit prices and dates of executions. These values should give at least a good approximation of what would have happened if the trading strategy had been used in the past. We call the part of a software system performing the historical simulation the “backtest engine”. It is essential that this part of the trading software works correctly in order to be able to rate the trading strategy. However, many trading platforms have an incorrect backtest engine, which can lead to serious misjudgments.

The following example illustrates this. A complete price chart contains the set of all prices determined at the stock market, and the corresponding points in time; each data point is called a tick. To obtain the chart, a huge amount of tick data is needed – in general, far more than necessary. Thus, a complete price chart is commonly compressed into so-called candle data. A candle is defined by the combination of the following four values used to summarize all price information about a given time period (in our example, we use a one-day period):

1. the maximum price during that time period (high);
2. the minimum price during the period (low);
3. the price at the beginning of the period (open); and
4. the price at the end of the period (close).

Now, assume that we only have daily candle data, ie, one candle per day, of Deutscher Aktienindex (DAX) futures. At the end of the day, we want to place a limit buy order, at the closing price of that day, which is valid for the next day. If an order is filled sometime during the following day, we immediately (intraday) place a profit target at the close of the last day (the limit buy trigger) plus €5. If we still hold the position at the end of the day, we update the profit target to the closing price of the current day plus €5. From Maier-Paape and Platen (2015), it is known that, when using candle data, the backtest cannot make a unique decision in each situation, ie, there are situations such that we cannot decide what would have happened (eg, in specific
candles first triggering the limit buy signal and then the target, or first triggering the target (nothing happens) and then triggering the buy signal). An alternative would be the use of tick data. This is still a finite set for any given time period, but other problems arise: there might be no tick data available for a time horizon of years or decades, and there is often far more data than is needed, which would dramatically increase computing time. Using candle data is still the de facto standard, despite this drawback and other limitations.

Nevertheless, even for undecidable situations, with candle data we can calculate the result in the best and worst cases. Using the results from Maier-Paape and Platen (2015), we can perform a backtest for both cases. We can also use real trading software, which we call “platform X” for anonymity. Results are shown in Figure 1 for DAX futures from 2000 until 2016, where we buy and hold at most one contract, and where one point of the DAX futures equals €1. From the results of platform X (upper solid curve), we believe that our trading strategy generates wins with only few and small drawdowns. In contrast, even the best case of a correct backtest evaluation (upper dashed curve) is not even half as good. In addition, the worst case (lower dashed curve) seems to be quite the opposite of the result of platform X. The correct answer for a backtest of this trading strategy would most likely be between the best and worst cases (both dashed curves). Hence, the results of platform X are unreliable. (Note that this says nothing about the live trading mode of platform X or any other trading software.)
This example shows that, even now, in 2017, there can be dangerous errors and shortcomings in widely used trading software that can misguide the user and lead to tremendous losses of real money.

There are a few more platforms that have been tested by the authors (see also Table 1): the software systems NinjaTrader and NanoTrader have some serious problems handling limit buy orders within the backtest feature. For example, NinjaTrader, in the version cited in Table 1, executes a limit buy order exactly at its limit buy trigger, even if the trigger is far above the stock price. This means that, even if the stock price is at €100 and the limit buy trigger at €2000, NinjaTrader will enter the position using €2000 as the entry price. On the other hand, NanoTrader, in its current version, has problems with limit buy orders coupled with a profit target. The platform Tradesignal works correctly but does not offer a choice of best, worst and ignore cases; the software itself decides which case to choose using its own logic. AgenaTrader has two backtest modes: the standard mode behaves like NinjaTrader; the advanced mode, which is the result of a collaboration by the AgenaTrader team and the authors, works correctly for basic order types and also provides best, worst and ignore cases in the sense of Maier-Paape and Platen (2015). However, there are more complex order types, such as if–then orders and one-cancels-other orders, which have not yet been tested within the trading platforms; even in AgenaTrader there could be errors.

The aim of this paper is to create scenarios for unit testing a backtest engine of commonly used trading tools, such that it is guaranteed that, under mild assumptions, each possible case is validated.

Note that, in addition to the above-mentioned problems, there are other limitations of backtesting that cannot be neglected (see, for example, Chan (2009, Chapter 3),

### TABLE 1  Comparison of backtest engines for some trading platforms and basic order types (ie, limit and stop orders).

<table>
<thead>
<tr>
<th>Platform</th>
<th>Company</th>
<th>Platform version</th>
<th>Do basic orders always work correctly for unique situations?</th>
<th>Backtest mode selectable (worst/best case)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>NanoTrader</td>
<td>Fipertec S.à.r.l.</td>
<td>3.2.0.81</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tradesignal</td>
<td>Tradesignal GmbH</td>
<td>7.7.2.35</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>NinjaTrader</td>
<td>NinjaTrader Group, LLC</td>
<td>7.0.1000.25</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>AgenaTrader</td>
<td>INCLUDE IT GmbH</td>
<td>1.8.0.441</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>(standard mode)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AgenaTrader</td>
<td>INCLUDE IT GmbH</td>
<td>1.8.0.441</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>(advanced mode)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pardo (2008, Chapter 6) and Harris (2008, Chapter 6); for trading options, see Izraylevich and Tsudikman (2012, Chapter 5). Since a backtest just simulates the behavior of a trading system over the past, it is strongly limited in predicting the future. However, it is common to optimize a parameter-dependent strategy on historical data to maximize some objective function. Such an optimization process can rapidly become very time consuming. Therefore, Ni and Zhang (2005) presented a method to efficiently backtest a trading system for different parameter choices, but they do not explain the backtest evaluation itself.

Computing such an “optimal” parameter setting does not ensure optimal parameters for the future, and can even lead to tremendous losses. Backtest overfitting is often referred to in this regard. This problem arises when more than one strategy configuration is tried, because the knowledge about the given historical data is always used in some sense to adjust the parameter. The more strategy configurations are tested, the more this strategy “learns” this specific historical data. The parameter setting leading to the highest performance on historical data may lead to underperformance on future data, because, for instance, the historical data mostly shows a bull market and in future there is a bear market. In this case, the optimization process may determine that an aggressive trading configuration is advantageous, which may fail in structurally different market phases. For a detailed discussion on backtest overfitting, see, for example, Bailey et al (2014, 2017), Carr and López de Prado (2014) and Pardo (1992, Chapter 6).

The above example shows that even a correct backtest engine needs to be applied carefully. Nevertheless, such an engine gives important information about a trading strategy and is the basis of any parameter optimization. The aim of this paper, however, is not the “right” application of backtest results, but how to verify whether or not backtests are performed correctly in the first place.

Common software solutions often struggle with the nonuniqueness of correct results of backtest engines (see the above example), and even situations that are obviously uniquely decidable (see Table 1). It is, however, of great importance for the users to have accurate backtest evaluations. Therefore, Maier-Paape and Platen (2015) asked how backtest engines should decide for certain given standard order setups, such as limit and stop entry orders, combined with typical intra-period stop or target exit orders, when only candle data is available. We could limit these examinations to single candles, arguing that any trade can be split up into several candles with different active orders. For the order combinations discussed, Maier-Paape and Platen provided decision trees with which the correct behavior of a backtest engine can be determined, for a given candle, depending on open, close, high and low values. They also introduced decision modes, which provide deterministic rules for situations that are not uniquely decidable due to a lack of information in the candle data. However, due to the numerous possibilities that can occur in different situations, it is hard to
verify whether or not a given backtest engine of some trading platform is calculating its backtests correctly. In this paper, we therefore provide the means for an algorithmic approach to this problem.

In order to check whether a given backtest engine behaves according to the requirements presented in the work of Maier-Paape and Platen (2015), it is necessary to design test data. This means that we are looking for a finite number of so-called “test candles”. After checking the backtest evaluation on these finitely many candles, we want to be able to conclude the backtest engine is generally correct under reasonable assumptions.

This paper is divided into five sections. In Section 2, we discuss the preliminaries under which our examinations are conducted and formalize the problem discussed. The main result of Section 3 is the development of the concept of “model candles”, which allows a “proof of correctness” of backtest engines in specific situations. In Section 4, a model for intra-period prices is introduced, which allows us to obtain the desired conditions that guarantee the completeness of all model candles and all corresponding possible outcomes of a backtest. Our conclusions are given in Section 5.

2 STATEMENT OF THE PROBLEM AND PRELIMINARIES

First, we specify the situation we shall examine. In Section 2.1, we state some assumptions needed to make the decisions for a backtest unique. Since we can reduce the problem of testing for correctness to a single candle for each situation, in Section 2.2 we define the setup for such situations. The concept of a backtest engine is explained in Section 2.3.

2.1 Assumptions

At this point, we will not make specific assumptions about the candles we examine. Clearly, every candle satisfies

\[ \text{low} \leq \min\{\text{open, close}\} \leq \max\{\text{open, close}\} \leq \text{high}. \]

Due to the lack of information caused by the candle data, some general assumptions must be made.

ASSUMPTIONS 2.1 (see Maier-Paape and Platen 2015, Section 2.1)

- No intra-period gaps: we assume that the intra-period price development during a candle is continuous.
- Market liquidity: all orders are filled at the requested price and thus without slippage.
- **Local worst case/best case:** worst case and best case decisions are determined locally, i.e., the profit of a trade is evaluated as if it were closed on the close of the candle.

- **Local decision:** the backtest engine decides by considering only the candle in question. Previous and future candles do not influence the decision made.

**Remark 2.2** Some of these assumptions are quite unreasonable for realistic price charts and markets. However, since the backtest engine has only candle data at hand, these assumptions are necessary and reasonable in deciding a possible outcome of these market situations.

To help clarify the concepts presented, we will use an example throughout this work.

**Example 2.3** At the beginning of a candle, we might have no open position but two active orders, e.g., a stop buy entry order at level 53, which enters a long position if the price reaches 53, and a protective stop loss order at level 51, which exits the long position (but only if it is opened beforehand) if the price falls below 51. Given an arbitrary candle, the backtest engine should now decide which of these orders is executed and whether or not there are open positions at the end of the candle.

### 2.2 The setup and intra-period prices

We can formalize which orders are active and what the position is prior to the candle.

**Definition 2.4** Suppose we have $m \in \mathbb{N}$ orders at levels $L_1 < \cdots < L_m$, respectively, with $L_i \in \mathbb{R}$ rounded to the tick size of the asset for all $1 \leq i \leq m$. Let $p \in \{-1, 0, 1\}$ denote the position status prior to the candle in question, where 0 stands for a flat position, $-1$ stands for a short position and 1 stands for a long position. A combination of $m$ such orders and the position type $p$ is called a “setup”.

In the following, we consider only setups that allow at most one entry execution and one exit execution. The available entry orders are limit buy/sell and stop buy/sell, i.e., “EnterLongLimit”, “EnterShortLimit”, “EnterLongStop” and “EnterShortStop”. As exit orders, we use stop loss and profit target (see Maier-Paape and Platen (2015, Sections 2.2, 2.3 and 2.5) for the relevant decision trees). Note that we neglect market orders for entry and exit, since they are trivially decidable. We do not investigate whether the results presented can be generalized to setups that allow at most $a$ entry and $a$ exit executions, where $1 < a \in \mathbb{N}$, but analogous concepts may be applicable. Further, we exclude the trivial case $m = 0$, in which no order is active.

Now, we can formalize Example 2.3 as follows.
Example 2.5 We use the combination of a stop buy entry (EnterLongStop) order at level $L_2 = 53$ and a protective stop loss exit order at level $L_1 = 51$ attached to that entry order, with $p = 0$ as an example setup.

For technical reasons only, the intra-period prices of a candle will be assumed to be continuous, where we denote the set of all continuous functions $g : A \to B$ by $\mathcal{C}(A,B)$ and define $\mathbb{R}^+ := \{x \in \mathbb{R} \mid x \geq 0\}$.

**Definition 2.6** We call $f \in \mathcal{C}([a,b], \mathbb{R}^+)$ with $a,b \in \mathbb{R}$ and $a < b$ an intra-period price function (IPF). We denote the corresponding candle of the IPF $f$ by

$$ C(f) := \left( f(a), f(b), \max_{t \in [a,b]} f(t), \min_{t \in [a,b]} f(t) \right) \in (\mathbb{R}^+)^4. $$

Here, $a$ and $b$ are interpreted as the start time and end time of a period that is summed up to a candle, and $f(t)$ as the price at time $t$ for $a \leq t \leq b$. Obviously, for each candle $c \in (\mathbb{R}^+)^4$ we can construct an IPF $f$ such that $C(f) = c$.

### 2.3 Backtest engines and results

Given an intra-period price function, we can define its result.

**Definition 2.7** We define the result of an IPF $f$ with respect to a given setup as the combination of $(\text{entry}_f, \text{exit}_f)$, where entry and exit result from the correct application of the orders in the given setup. If entry or exit do not occur, they are denoted by the value $-1$. We denote the result of the IPF $f$ by $R(f)$. Formally, for a given setup, this can be regarded as the function

$$ R : \bigcup_{a,b \in \mathbb{R}, a < b} \mathcal{C}([a,b], \mathbb{R}^+) \to (\mathbb{R}^+ \cup \{-1\})^2. \quad (2.1) $$

With this, the combined candle data and result (CR) of $f$ with respect to a given setup can be defined as the combination $\text{CR}(f) := (C(f), R(f))$.

**Remarks 2.8** Note that results and CRs are defined depending on IPFs, which means that all CRs correspond to intra-period prices that fulfill the first two of Assumptions 2.1.

For given setup and candle data (open, close, high, low), finding all possible results (entry, exit) can be reformulated as finding $\text{im}(\text{CR}) \cap ((\text{open, close, high, low}) \times \mathbb{R}^2)$, where $\text{im}(\text{CR})$ denotes the image of the map $\text{CR}$. Note that this set can have more than just one element, which means that the result for this candle is not unique.

For a given fixed setup, we can formalize the concept of a backtest engine used in this work. As we want to examine the behavior of a backtest engine on candle data, we can take such a candle as input data. The backtest engine should return a possible
entry price at which a position may be opened and a possible exit price at which the position may be closed. Of course, entry or exit might not occur or might not have a price that can be uniquely determined (see Remark 2.8). Thus, the backtest engine also needs a “backtest decision mode” (backtest mode, BM), which makes the entry and exit prices unique.

**Definition 2.9** For BM := \{best case, worst case, ignore\}, we interpret a value in \((\mathbb{R}^+)^4 \times \text{BM}\) as the combination of the candle data (open, close, high, low) \(\in (\mathbb{R}^+)^4\) and the backtest mode according to Assumptions 2.1.

For a given setup, we call a mapping

\[ E : (\mathbb{R}^+)^4 \times \text{BM} \rightarrow (\mathbb{R}^+ \cup \{-1\})^2 \]

a “backtest engine”.

The result \(E(c_0, \mathfrak{M}) \in (\mathbb{R}^+ \cup \{-1\})^2\) for a candle \(c_0 \in (\mathbb{R}^+)^4\) and a backtest mode \(\mathfrak{M} \in \text{BM}\) is interpreted as the combination of entry price and exit price, where again \(-1\) denotes no entry/no exit.

**Remark 2.10** On a particular chart, here a candlestick chart, the backtest engine of a platform decides the outcomes of a given strategy. It has to execute the (user-defined) code of the trading approach that is typically called at the end of each candle and places orders. Further, it has to keep track of placed orders and open positions. When all orders placed before a candle, which form the setup, are known, the backtest engine has to decide which of these orders are executed. Here, we consider only this last part of the backtest engine. So, for us, a backtest engine \(E\) has only the candle data (and the setup) as input, whereas the result function \(R\) from (2.1) requires an IPF, ie, intra-period data. Thus, \(E\) has to find results that correspond to the given candle, ie, for a given candle \(c_0\), the backtest engine has to find a result \(r_0 \in (\mathbb{R}^+ \cup \{-1\})^2\) such that \(r_0 = R(f)\) and \(c_0 = C(f)\) for an appropriate IPF \(f\). The backtest mode specifies which IPF has to be chosen if there are several possible results corresponding to the given candle.

Continuing Example 2.5, we show these concepts in an application.

**Example 2.11** Using the setup from Example 2.5 (stop buy entry order at \(L_2 = 53\), stop loss exit order at \(L_1 = 51\) with a flat position before the candle, ie, \(p = 0\)) and the IPF

\[ f : [0, \frac{5}{2}\pi] \rightarrow \mathbb{R}^+ , \quad t \mapsto \sin(t) + 52 \]

(see Figure 2(a)), we obtain the CR

\[
(C(f), R(f)) = (\text{open} = 52, \text{close} = 53, \text{high} = 53, \text{low} = 51, \text{entry}_f = 53, \text{exit}_f = 51).
\]
In this example, we can see the nonuniqueness of results, if only a candle is given, by considering the IPF
\[ g : [0, \frac{3}{2} \pi] \rightarrow \mathbb{R}^+, \quad t \mapsto -\sin(t) + 52 \] (see Figure 2(b)), with the same resulting candle as for \( f \), ie, \( C(f) = C(g) \), but with the result
\[ R(g) = (\text{entry}_g = 53, \text{exit}_g = -1) \neq R(f). \]
Here, \( g \) corresponds to the best case for this candle, and \( f \) to the worst case, as, with a fictional sale at the close of the candle, we have for the cash value of \( f \) and \( g \) that
\[
\text{close} - \text{entry}_g = \text{close} - 53 = 53 - 53 = 0 > -2 = 51 - 53 = \text{exit}_f - \text{entry}_f.
\]
At this point, we actually do not know whether there could be other cases, but for this setup there are at most two possible results for a given candle. Choosing a backtest mode \( \mathcal{M} \in \text{BM} \) then makes the decision for the backtest engine unique and allows us to choose the result from either \( f \) or \( g \) or to ignore this trade.

Since we cannot check a backtest engine for each candle contained in the infinite-dimensional space \((\mathbb{R}^+)^4\), we need to reduce the problem to finitely many model candles. However, for this step we need the backtest engine to be stable under the transformation of candles, which we discuss in the next section.

### 3 SUITABLE TEST CASES FOR A PROOF OF CORRECTNESS

In this section, we want to examine whether or not a given backtest engine works correctly. For this goal, we design “model candles”, ie, a set of finitely many candles that allows a proof of correctness under the assumption of “stability under transformations”, a concept that will be introduced here.

In general, the values of a CR do not coincide with the levels of the underlying setup, but we observe that possible values for entry and exit prices are the “open”
of the candle and the levels of the setup. The exact values of open, close, high and low are not important for the resulting entry and exit, but their position relative to the levels $L_1, \ldots, L_m$ is. We therefore define $l_{2i-1} := L_i$ for $i = 1, \ldots, m$ and introduce intermediate levels $l_{2i}$ with $L_i < l_{2i} < L_{i+1}$ such that $l_0 < l_1 < \cdots < l_{2m}$. Restricting the candle data to these values results in a system of finitely many representative candles. We observe that the number of representative candles is at most $(2m + 1)^4$, as open, close, high and low can only take the $2m + 1$ values $l_0, \ldots, l_{2m}$.

**Example 3.1** In Example 2.5, we may choose

$$l_0 := 50 < l_1 = L_1 = 51 < l_2 := 52 < l_3 = L_2 = 53 < l_4 := 54.$$  \hspace{1cm} (3.1)

Figure 3 shows all representative candles for this setup.

In order to further restrict the levels $l_i$ to fixed values, we need to introduce the notion of transformation from one set of levels to another.

**Definition 3.2** In the following, a strictly monotonously increasing bijective function

$$T : \mathbb{R}^+ \rightarrow \mathbb{R}^+$$

is called a transformation. In order to apply a transformation to our results, we set $T(-1) := -1$ for entries and exits. A transformation can be used to transfer setups, candles, IPFs, results and CRs from one set of levels to another.

**Example 3.3** Continuing Example 3.1 with the levels given in (3.1) and examining the corresponding CR $(52, 51, 53, 51, 53, 51)$ of the IPF $f$ from Example 2.11, we might apply the transformation

$$T : \mathbb{R}^+ \rightarrow \mathbb{R}^+, \quad t \mapsto 2t + 1.$$

This results in a setup with the same orders but at levels

$$\tilde{l}_0 := 101 < \tilde{l}_1 = \tilde{L}_1 = 103 < \tilde{l}_2 := 105 < \tilde{l}_3 = \tilde{L}_2 = 107 < \tilde{l}_4 := 109$$

and a transformation of the CR given by $(105, 103, 107, 103, 107, 103) \in \mathbb{R}^6$. 
FIGURE 4 An example of candles with values in \( \{l_{2,j} \mid j = 1, \ldots, 4\} \) (see Definition 3.6): (a) “model candles” and (b) corresponding “representative candle”.

\[
\begin{align*}
L_2 &= l_3 \\
L_2 &= l_3
\end{align*}
\]

Apparently, this tuple is itself a CR, ie, it is derived from an IPF. This can be shown by using the IPF \( f \) from Example 2.11 and transforming it to \( T \circ f \).

Now we introduce the main assumption about the given backtest engine, which will allow us to prove correctness of that backtest engine.

**Definition 3.4** We call a backtest engine \( E \) stable under transformations if the following holds for any given IPF \( f \), transformation \( T \) and setup at levels \( L_1, \ldots, L_m \):

- if the result of the backtest engine \( E \) with backtest mode \( \mathfrak{M} \in \text{BM} \) is given by
  \[
  E(C(f), \mathfrak{M}) = (\text{entry}, \text{exit}),
  \]
  then the result of the backtest engine after transformation for the new setup at levels \( T(L_1), \ldots, T(L_m) \) and the new IPF given by \( T \circ f \) yields
  \[
  E(C(T \circ f), \mathfrak{M}) = (T(\text{entry}), T(\text{exit})) =: T(\text{entry}, \text{exit}).
  \]

**Remark 3.5** For a backtest engine, it is desirable to be stable under transformations. This is due to the structure of orders that only ask for the relative position between some trigger level and the price, ie, whether the price is above or below that level; the exact values of the trigger and price are irrelevant. The decision trees in the work of Maier-Paape and Platen (2015), which outline how correct results for several order setups can be obtained, are also independent of the exact levels and only depend on relative values. Thus, both the decision tree and the corresponding results are naturally stable under transformations. We conclude that the results depend only on the relative positions of orders, or, formally, for all IPFs \( f \) we have \( R(T \circ f) = T(R(f)) \).
Additionally, by the monotonicity of transformations, best cases remain best cases and worst cases remain worst cases under transformations, which can again be seen in the decision trees of Maier-Paape and Platen (2015).

From representative candles with values in \( \{l_0, \ldots, l_{2m}\} \), we can now cover almost all possible candles by applying transformations, but some important cases are missing: a “generic” candle has a range \([\text{low}, \text{high}]\) that does not include any level from \(\{L_1, \ldots, L_m\}\), but nonetheless the values for open, close, high and low may differ from each other (see Figure 4(a) for examples of such candles). These cases are not covered by the considerations made so far, but can easily be taken into account by the following considerations. Combining our observations, we arrive at the set of “model” candles we want to examine by introducing four intermediate levels between \(L_{i-1}\) and \(L_i\).

**Definition 3.6** Let a setup at levels \(L_1, \ldots, L_m\) be given. Further, choose values \(l_{2i}, l_{2i-1}, l_{2i,j} \in \mathbb{R}^+\) for \(i = 1, \ldots, m\) and \(j = 1, \ldots, 4\) with

\[
l_0 = l_{0,1} < l_{0,2} < l_{0,3} < l_{0,4} < L_0, \\
l_{2i-1} := L_i < l_{2i} = l_{2i,1} < l_{2i,2} < l_{2i,3} < l_{2i,4} < l_{2i+1} := L_{i+1} \\
\text{for } 1 \leq i \leq m - 1, \\
L_m < l_{2m} = l_{2m,1} < l_{2m,2} < l_{2m,3} < l_{2m,4}.
\]

A model candle is a candle with values

\(\text{open, close, high, low} \in \{l_i \mid 0 \leq i \leq 2m\} \cup \{l_{2i,j} \mid 0 \leq i \leq m, 1 \leq j \leq 4\}\),

with the further restriction

\[
\forall i \in \{0, \ldots, m\} \exists r_i \in \{1, \ldots, 4\}: \\
\text{if } \{\text{open, close, high, low}\} \cap \{l_{2i,j} \mid 1 \leq j \leq 4\} \neq \emptyset, \\
\text{then } \{\text{open, close, high, low}\} \cap \{l_{2i,j} \mid 1 \leq j \leq 4\} = \{l_{2i,j} \mid 1 \leq j \leq r_i\}. \tag{3.2}
\]

Figure 4 shows examples of some candles where, in this extremal case, we have up to twelve model candles for one of the representative candles introduced.

In this way, we obtain sufficiently many candles to cover the “generic” candles described above as well.

**Remarks 3.7** Condition (3.2) for the sublevels \(l_{2i,j}\) is introduced to minimize the use of these additional levels by, figuratively speaking, filling up the levels \(l_{2i,j}\) for a fixed \(i\) by increasing \(j\) as far as necessary. This reduces the total number of model candles.

For an application of the construction, all values can be chosen to be equidistant. It is important to consider the tick size used and to not choose the sublevels \(l_{2i,j}\) with
too small a distance. For example, if we have a tick size of 0.01, we could choose \( l_i := 50.05 + i \) and \( l_{2i,j} := 49.95 + 2i + 0.1j \) for \( i = 0, \ldots, m \), \( j = 1, \ldots, 4 \), which also allows us to check for rounding errors.

With all these preparations in order, we are now able to conduct a proof of correctness for given backtest engine and setup.

**Theorem 3.8** Let a backtest engine \( E \) that is stable under transformations for a setup at (ordered) levels \( L_1 < \cdots < L_m \) be given. If the backtest engine works correctly (eg, for the criteria given by Maier-Paape and Platen (2015)) on the set of all model candles of this setup with fixed \( l_i, l_{2i,j} \), as in Definition 3.6, then it works correctly on this setup with any (ordered) levels \( \tilde{L}_1 < \cdots < \tilde{L}_m \) (and arbitrary candles).

**Proof** Let an arbitrary candle \( c = (\text{open, close, high, low}) \in (\mathbb{R}^+)^4 \), ordered levels \( \tilde{L}_1, \ldots, \tilde{L}_m \in \mathbb{R}^+ \) and a backtest mode \( M \) be given. Set \((\text{entry, exit}) := E(c, M)\).

We need to show that \((\text{entry, exit})\) is a correct result for the candle \( c \), ie, that there exists some IPF \( \tilde{f} \) with \( c = C(\tilde{f}) \) and \( R(\tilde{f}) = (\text{entry, exit}) \), or, if \( M = \) ignore and there is no unique result, that \((\text{entry, exit}) = (-1, -1)\).

In order to do this, let us define \( \tilde{l}_{2i-1} := \tilde{L}_i \) for \( 1 \leq i \leq m \) and \( \tilde{l}_{2i,j}, \tilde{L}_i \in \mathbb{R}^+ \) such that the requirements from Definition 3.6 are fulfilled and open, close, high, low \( \in \{\tilde{L}_i \mid 0 \leq i \leq 2m\} \cup \{\tilde{l}_{2i,j} \mid 0 \leq i \leq m, 1 \leq j \leq 4\} \), which fulfill the minimal sublevel condition (3.2) for the level \( \tilde{l}_{2i,j} \) instead of \( l_{2i,j} \). Such a choice is obviously always possible.

Now, we construct a transformation \( T \) that transforms the setup at levels \( \tilde{L}_i \) to the setup at levels \( L_i \). This can be done by setting \( T(0) = 0 \), \( T(\tilde{l}_{2i-1}) = l_{2i-1} \), \( T(\tilde{l}_{2i,j}) = l_{2i,j} \) for \( i = 0, \ldots, m \) and \( j = 1, \ldots, 4 \) and interpolating piecewise linearly in between these values. Formally,

\[
T : \mathbb{R}^+ \to \mathbb{R}^+,
\]

\[
t \mapsto \begin{cases}
\frac{l_{0,1}}{0,1}, & t < \tilde{l}_{0,1}, \\
\frac{l_{2i,j} + (t - \tilde{l}_{2i,j})}{l_{2i,j+1} - l_{2i,j}} & 0 \leq i \leq m, 1 \leq j \leq 3, \\
\frac{l_i + (t - \tilde{l}_i)}{l_{i+1,1} - l_i}, & \tilde{l}_i \leq t < \tilde{l}_{i+1,1}, i \text{ odd}, \\
\frac{l_{i-1,4} + (t - \tilde{l}_{i-1,4})}{l_i - l_{i-1,4}}, & \tilde{l}_{i-1,4} \leq t < \tilde{l}_i, i \text{ odd}, \\
\tilde{l}_{2m,4} - \tilde{l}_{2m,4} + t, & t \geq \tilde{l}_{2m,4}.
\end{cases}
\]
FIGURE 5 An example of the application of the transformation $T$ on an arbitrary candle.

(a) $L_{i+1} = \widetilde{l}_{2i+1}$ \hspace{1cm} (b) $L_{i+1} = \tilde{l}_{2i+1}$

$\widetilde{l}_{2i} = \widetilde{l}_{2i,1}$ \hspace{1cm} $l_{2i,4}$

$\widetilde{l}_{2i} = \widetilde{l}_{2i,1}$ \hspace{1cm} $l_{2i}$

$\widetilde{l}_{2i} = \widetilde{l}_{2i-1}$ \hspace{1cm} $L_i = l_{2i-1}$

(a) Candle $c$ and levels $\widetilde{L}_i, \widetilde{l}_i, \widetilde{l}_{i,j}$. (b) Candle $T(c)$, a model candle with respect to $L_i, l_i, l_{i,j}$.

By construction, $T$ is continuous, strictly monotonously increasing and bijective. Further, $T(c) := (T(\text{open}), T(\text{close}), T(\text{high}), T(\text{low}))$ is a model candle with respect to $L_i, l_i, l_{i,j}$ (see Figure 5).

By assumption, $E$ works correctly on model candles with respect to $L_i, l_i, l_{i,j}$. Therefore, we know that $E(T(c), \mathcal{M})$ is the desired result of the model candle $T(c)$ and $(T(c), E(T(c), \mathcal{M}))$ is the desired CR.

If $\mathcal{M} = \text{ignore}$ and there is no unique result for $T(c)$, this still holds after applying the inverse transformation $T^{-1}$, so we have $(-1, -1) = E(T(c), \mathcal{M}) = E(c, \mathcal{M})$ by stability under transformations, which is the correct result by Remark 3.5.

Otherwise (ie, $\mathcal{M} \neq \text{ignore}$, or the result for $T(c)$ is unique), there exists an IPF $f = f_{T(c), \mathcal{M}}$ (depending on $T(c)$ and $\mathcal{M}$ only) such that $C(f) = T(c)$ and $R(f) = E(T(c), \mathcal{M})$, because each result corresponds to some IPF. Then, clearly, $C(T^{-1} \circ f) = c$. Therefore, setting $\tilde{f} := T^{-1} \circ f$, we already have $C(\tilde{f}) = c$, as desired. Further, applying the stability of $E$ under the transformation $T^{-1}$ (see Definition 3.4), we obtain

$$(\text{entry}, \text{exit}) \overset{\text{def.}}{=} E(c, \mathcal{M}) = E(C(T^{-1} \circ f), \mathcal{M}) \overset{\text{stability u.t. of } E}{=} E(T^{-1}(E(C(f), \mathcal{M}))).$$

Since, by construction, $C(f) = T(c)$ and $R(f) = E(T(c), \mathcal{M})$, we conclude from the stability under transformation (u.t.) of results (see Remark 3.5) that

$$(\text{entry}, \text{exit}) = T^{-1}(E(T(c), \mathcal{M})) = T^{-1}(R(f)) = R(\tilde{f}),$$

which is, again by Remark 3.5, the correct result.

As $c$ and $\widetilde{L}_1, \ldots, \widetilde{L}_m$ were chosen arbitrarily, we conclude general correctness of $E$ for this setup. \hfill $\square$
Theorem 3.8 allows us to reduce the verification of a stable backtest engine to simply checking the backtest evaluation on a finite set of model candles. Therefore, this set is fixed in the next section.

4 A MODEL FOR INTRA-PERIOD PRICES

Our goal in this section is to develop a model for IPFs that allows only finitely many model price functions but still results in all possible CRs. The results for all CRs are needed to test a given backtest engine by comparing its results with correct results. They can be determined by the means of this section with the help of a correct reference backtest engine on the finitely many model price functions.

We may restrict ourselves to model candles that do not take the values \( l_{2i,2}, l_{2i,3}, l_{2i,4} \), but only \( l_{2i} \) for all \( 1 \leq i \leq m \), because, as we see below, we can reconstruct the results for the model candles with values in \( l_{2i,2}, l_{2i,3}, l_{2i,4} \).

Remark 4.1 Let a setup and a backtest mode \( \mathcal{R} \) be given.

Further, let \( c = (\text{open}, \text{close}, \text{high}, \text{low}) \) be a model candle with respect to the levels

\[
\mathcal{L} := \{l_i \mid 0 \leq i \leq 2m\} \cup \{l_{2i,j} \mid 0 \leq i \leq m, 1 \leq j \leq 4\},
\]

i.e., \( c \in \mathcal{L}^4 \). We can obtain the result for \( c \) by first applying the substitution

\[
l_{2i,j} \mapsto l_{2i} \quad \forall 0 \leq i \leq m, 1 \leq j \leq 4,
\]

to \( c \), which leads to a new candle, \( c' \). Calculating the result \( r' \) of \( c' \) and then “re-substituting”

\[
l_{2i_0} \mapsto l_{2i_0,j_0} \quad \text{in case } l_{2i_0,j_0} = \text{open of the candle } c
\]

in \( r' \) gives us the result \( r \). Indeed, \( r \) is the desired result for \( c \).

This procedure is possible due to the fact that entry and exit can only occur at levels \( L_i \) or at the open value.

Example 4.2 Remark 4.1 can be illustrated for our setup from Example 2.5 (EnterLongStop at \( L_2 \), stop loss at \( L_1 \)) with the model candle

\[
c := (\text{open} = l_{4,2}, \text{close} = l_{4,1} = l_4, \text{high} = l_{4,3}, \text{low} = l_3 = L_2),
\]

which has the unique result \( r := (\text{entry} = \text{open} = l_{4,2}, \text{exit} = -1) \). By substituting \( l_{4,j} \) with \( l_4 \) for all \( 1 \leq j \leq 4 \), we get the representative candle

\[
c' := (\text{open}' = l_4, \text{close}' = l_4, \text{high}' = l_4, \text{low}' = l_3 = L_2)
\]

with the unique result \( r' := (\text{entry}' = \text{open}' = l_{4,2}, \text{exit}' = -1) \). We can reconstruct \( r \) from \( r' \) by substituting \( l_4 = \text{open}' \leftrightarrow \text{open} = l_{4,2} \) in \( r' \).
The following model for intra-period prices allows only the values $l_0, \ldots, l_{2m}$ at certain interpolation points.

**Definition 4.3** An intra-period model price series (IPMS) with respect to a given setup is a finite sequence $s \triangleq (l_{i_1}, \ldots, l_{i_k})$ with $0 \leq i_j \leq 2m$ for all $1 \leq j \leq k$ and $i_{j+1} = i_j + 1$ or $i_{j+1} = i_j - 1$ for all $1 \leq j \leq k - 1$. We call $|s| \triangleq k$ the size of the IPMS. We denote by $\mathcal{M} \triangleq \mathcal{M}_m$ the set of all IPMSs with respect to a number of levels, $m$.

The connection with intra-period price functions is the following.

**Observation 4.4** Every IPMS $s = (l_{i_1}, \ldots, l_{i_k})$ can be interpreted as an IPF by using piecewise linear interpolation as follows. Define $f^{(s)} : [1, k] \to \mathbb{R}^+$ by

$$f^{(s)}(t) := l_{i_{\lfloor t \rfloor}} + (t - \lfloor t \rfloor)(l_{i_{\lfloor t \rfloor} + 1} - l_{i_{\lfloor t \rfloor}}), \quad \text{where } l_{i_{k+1}} := l_{i_k},$$

such that $f^{(s)}(j) = l_{i_j}$ for $j = 1, \ldots, k$. Indeed, $f^{(s)}$ is continuous by construction.

Using this connection, we define the result of an IPMS as follows.

**Definition 4.5** For a given setup with $m$ levels, we define

$$R : \mathcal{M} \to (\mathbb{R}^+ \cup \{-1\})^2, \quad s \mapsto R(f^{(s)}),$$

and call $R(s)$ the result of $s$.

Similarly, we denote by $C(s) \triangleq C(f^{(s)})$ the candle resulting from $s$.

**Example 4.6** In Example 2.5, an IPMS may be given by $s = (l_2 = 52, l_3 = 53, l_2 = 52, l_1 = 51, l_2 = 52, l_3 = 53)$. A graphical representation of the corresponding piecewise linear IPF $f^{(s)}$ is given in Figure 6.
This example has the same CR as the IPF $f$ in Example 2.11, which is generalized in the following theorem.

**Theorem 4.7** Every $\text{CR } r \in \{l_0, \ldots, l_{2m}\}^4 \times (\mathbb{R}^+ \cup \{-1\})^2$ is the CR of an IPMS.

**Proof** Let $f \in C([a, b], \mathbb{R}^+)$ be an IPF with $C(f) \in \{l_0, \ldots, l_{2m}\}^4$ such that $(C(f), R(f)) = r$ for an arbitrary but fixed CR $r$. Such an $f$ exists by the definition of CRs (see Definition 2.7).

In order to construct an IPMS with the same result $r$, we use piecewise linear interpolation: without loss of generality, we assume that $f$ takes values in $\{l_0, \ldots, l_{2m}\}$ only at finitely many discrete points, because otherwise we can easily modify $f$ in such a way without changing the CR $r$. Now define $t_1, \ldots, t_k$ such that

$$t_j < t_{j+1} \quad \text{for all } 1 \leq j \leq k-1,$$

$$f(t_j) \in \{l_0, \ldots, l_{2m}\} \quad \text{for all } 1 \leq j \leq k,$$

and

$$f(t) \notin \{l_0, \ldots, l_{2m}\} \quad \text{for all } t \in [a, b] \setminus \{t_1, \ldots, t_k\}.$$

Next, we remove all those $t_j$ for which $f(t_j) = f(t_{j-1})$ from the $\{t_1, \ldots, t_k\}$ (remember that one requirement for an IPMS $s$ is that $s_{j+1} \neq s_j$) by defining $t'_j$ such that

$$\{t'_1, \ldots, t'_{k'}\} = \{t_1, \ldots, t_k\} \setminus \{t_j \mid f(t_j) = f(t_{j-1})\}, \quad t'_j < t'_{j+1} \quad \forall 1 \leq j \leq k'-1.$$

By continuity of $f$ and the intermediate value theorem, $s := (f(t'_1), \ldots, f(t'_{k'}))$ is an IPMS, and, again by the intermediate value theorem, $R(s) = R(f) = r$. \qed

Example 4.6 in fact gives the IPMS obtained from $f$ in Example 2.11 by the construction in the proof of Theorem 4.7.

For an algorithmic approach, we want to limit ourselves to finitely many IPMSs from which all CRs can be obtained. For this, we need to limit the size of the IPMSs.

**Definition 4.8** We denote by $M_n$ the set of all CRs of IPMSs of size of at most $n$, i.e.,

$$M_n := \{r \mid \exists s \in \mathcal{M}, |s| \leq n, r = (C(s), R(s))\}.$$

Now we can limit the size of IPMSs used to obtain all possible CRs for a given setup.

**Theorem 4.9** For a given setup and $n_0 \in \mathbb{N}$, the following holds: if $M_{n_0} = M_{n_0+1}$, then $M_n = M_{n_0}$ for all $n \geq n_0$. 

---

*To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net*
**FIGURE 7** Construction of smaller IPMS with the same CR (see Proof of Theorem 4.9 and Example 4.12).

(a) Initial IPMS $s$. (b) IPMS $s'$. (c) IPMS $\tilde{s}'$ with the same result as $s'$, but smaller size. (d) IPMS $\tilde{s}$ with the same result as $s$, but smaller size.

**Proof** We show that $M_n = M_{n+1}$ implies $M_{n+1} = M_{n+2}$. Then, the claim follows inductively.

There are four steps in the following construction. These steps are illustrated by Figure 7.

1. Let $M_n = M_{n+1}$ and $s = (l_{i_1}, \ldots, l_{i_{n+2}})$ be an IPMS of size $n + 2$ with CR $r = (\mathcal{C}(s), \mathcal{R}(s)) = (\text{open, close, high, low, entry, exit}) \in M_{n+2}$ (see Figure 7(a)). We have to construct an IPMS of size of at most $n + 1$ that has the same CR, $r$.

2. Consider the IPMS $s' = (l_{i_1}, \ldots, l_{i_{n+1}})$, which is the same IPMS as $s$ but shortened by the last element, i.e., we have $s = (s', l_{i_{n+2}})$ (see Figure 7(b)). The size of $s'$ is $n + 1$ and we denote its CR by $r' = (\mathcal{C}(s'), \mathcal{R}(s')) = (\text{open', close', high', low', entry', exit'}) \in M_{n+1}$.

3. By the definition of an IPMS we have $|\text{close'} - l_{i_{n+2}}| = 1$. By assumption we have $r' \in M_n$, i.e., there exists an IPMS $\tilde{s}'$ of size $|\tilde{s}'| \leq n$ with CR $r'$ (see Figure 7(c)). Without loss of generality let $|\tilde{s}'| = n$ (otherwise, the same proof can be conducted with a different index in use).

4. We construct $\tilde{s} \in \mathbb{R}^{n+1}$ as the concatenation $\tilde{s} := (\tilde{s}', l_{i_{n+2}})$ with CR $\tilde{r} = (\mathcal{C}(\tilde{s}), \mathcal{R}(\tilde{s})) = (\text{open, close, high, low, entry, exit})$ (see Figure 7(d)). Indeed, $\tilde{s}$ is an IPMS, as $\tilde{s}_n = \tilde{s}'_n = \text{close'}$, and, by definition, $|\tilde{s}_n - \tilde{s}_{n+1}| = |\text{close'} - l_{i_{n+2}}| = 1$. 

www.risk.net/journals

To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
We now claim that $\bar{s}$ has the CR $\bar{r}$, i.e., $\bar{r} = r$. By viewing the structure as a concatenation, it is obvious that $\overline{\text{open}}' = \text{open}' = \text{open}$ and $\overline{\text{close}} = l_{n+2} = \text{close}$. Further, the high and low values can be determined as the maximum and minimum, respectively:

$$\begin{align*}
\overline{\text{high}} &= \max\{\max \bar{s}', l_{n+2}\} \overset{e(s') = e(s')}{=} \max\{\text{high}', l_{n+2}\} \\
\overline{\text{low}} &= \min\{\text{low}', l_{n+2}\} \overset{e(s') = e(s')}{=} \min\{\text{low}', l_{n+2}\}
\end{align*}$$

Therefore, we have $C(\bar{s}) = C(s)$.

If there is no entry for $s$, the same holds for $\bar{s}$ because the range $[\text{low}, \overline{\text{high}}]$ contains no entry order level. If the entry occurs during the first $n + 1$ interpolation points of $s$, we get $\overline{\text{entry}} = \text{entry}' = \text{entry}$. Otherwise, the entry of $s$ is executed on the $(n + 2)$th interpolation point with value $\text{entry} = l_{n+2}$. Then, $\text{entry}' = \overline{\text{entry}} = l_{n+2}$. Therefore, $\text{entry} = \overline{\text{entry}}$. Analogously, it follows that exit = exit.

In conclusion, $R(\bar{s}) = R(s)$, and thus $\bar{r} = r$, but $|\bar{s}| = n + 1$ and therefore $r \in M_{n+1}$.

This allows an algorithmic approach to find all relevant CRs.

**Corollary 4.10** In order to obtain all possible CRs of model candles with values in $\{l_0, \ldots, l_{2m}\}$ for a given setup, it suffices to compute $M_{n_0}$, where $n_0 = \min\{n \mid M_n = M_{n+1}\}$.

**Proof** The claim follows from Theorems 4.7 and 4.9.

**Remark 4.11** In the proof of Theorem 4.9, we showed that for $\bar{s}'$ and $s'$ the equality

$$(C(\bar{s}'), R(\bar{s}')) = (C(s'), R(s'))$$

carries forward after appending $l_{n+2}$ as

$$(C(\bar{s}', l_{n+2}), R(\bar{s}', l_{n+2})) = (C(s', l_{n+2}), R(s', l_{n+2})).$$

Analogously, and by induction, we obtain the following: if the CRs of two IPMSs $s_1$ and $s_2$ are the same, then so are the CRs of all IPMSs obtained from $s_1$ and $s_2$ by concatenation of arbitrary consistent IPMSs, i.e., if $C(s_1) = C(s_2)$ and $R(s_1) = R(s_2)$, then

$$\forall k \in \mathbb{N}, \ s_3 \in \mathbb{R}^k \text{ such that } (s_1, s_3) \text{ and } (s_2, s_3) \text{ are IPMSs,}$$

$$C((s_1, s_3)) = C((s_2, s_3)) \quad \text{and} \quad R((s_1, s_3)) = R((s_2, s_3)).$$

Therefore, an algorithm only needs to consider one of these IPMSs when appending other IPMSs, which speeds up the calculation of all model candles and their results enormously.
Example 4.12  Continuing Example 2.5, we can apply the results. To abbreviate our notation, we set $l_i := i$. Algorithmically, we find the value $n_0 = 11$ to be sufficiently large in the sense of Corollary 4.10. The construction from the proof of Theorem 4.9 shown in the examples in Figure 7 has the following data:

$$s = (1, 2, 3, 4, 3, 2, 1, 0, 1, 2, 3, 4, 3),$$

$$|s| = 13,$$

$$(C(s), R(s)) = (\text{open} = 1, \text{close} = 3, \text{high} = 4, \text{low} = 0, \text{entry} = 3, \text{exit} = 1),$$

$$s' = (1, 2, 3, 4, 3, 2, 1, 0, 1, 2, 3, 4),$$

$$|s'| = 12,$$

$$(C(s'), R(s')) = (\text{open}' = 1, \text{close}' = 4, \text{high}' = 4, \text{low}' = 0, \text{entry}' = 3, \text{exit}' = 1),$$

$$\bar{s}' = (1, 2, 3, 2, 1, 0, 1, 2, 3, 4),$$

$$|\bar{s}'| = 10,$$

$$(C(\bar{s}'), R(\bar{s}')) = (C(s'), R(s')),$$

$$\bar{s} = (1, 2, 3, 2, 1, 0, 1, 2, 3, 4, 3),$$

$$|\bar{s}| = 11,$$

$$(C(\bar{s}), R(\bar{s})) = (C(s), R(s)).$$

5 CONCLUSION

In this paper, we provided tools to test the correctness of backtest engines for setups with at most one entry and one exit. Many practical situations are covered by these tools, such as “EnterLongLimit”, “EnterShortLimit”, “EnterLongStop” and “EnterShortStop” with accompanying intra-period stop loss and target orders.

By constructing all relevant intra-period price functions, which we could limit to finitely many intra-period model price series (IPMSs), and then running a reference backtest for IPMSs, we obtained the correct result for all model candles. Comparing these results with the results of a given backtest engine on all model candles, we could decide on the correctness of the backtest engine under an assumption of stability under transformations by the results of Section 3.

Many of those concepts can be generalized to more complex situations with more than one entry or exit. It remains to be shown how our results can be transferred to these setups. Further, in these cases there would not necessarily be unique worst and best cases anymore.

Another extension of our work could be the consideration of other order types, eg, one-cancels-other orders, which would require a revision of the theory presented.

www.risk.net/journals  Journal of Investment Strategies
To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net
DECLARATION OF INTEREST

R. Löw was funded by Undergraduate Fund Project 2015, RWTH Aachen. The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES


Research Paper

Statistical testing of DeMark technical indicators on commodity futures

Marco Lissandrin, Donnacha Daly and Didier Sornette

Chair of Entrepreneurial Risks, ETH Zürich, Scheuchzerstrasse 7, SEC F 7, CH-8092 Zürich, Switzerland; emails: lissandrin.marco@gmail.com, donnacha@ieee.org, dsornette@ethz.ch

(Received November 29, 2016; revised April 24, 2017; accepted May 11, 2017)

ABSTRACT

In this paper, we examine the performance of three DeMark indicators (Sequential, Combo and Setup trend), which constitute specific implementations of technical analysis often used by practitioners, over twenty-one commodity futures markets and ten years of daily data. Our work addresses price behavior following new entry signals by studying whether, for short holding periods, the entry signals generated by these indicators can time the market moves and suggest the right side of the market (long or short). For example, we want to know how long we should hold or delay a trade before the price is expected to move significantly. The signals are sparse, as they mostly suggest entering the market between one and five times per year. To adjust for the limited number of total days in which the trade signals are in-the-market, we generate the distributions of multiple performance metrics (mean return, profit factor and risk–return ratio) over different trade holding horizons, and compare them with their randomized versions, which have the same number of entry signals and the same number of holding days. The rolling strategy, which creates continuous futures data from separate contracts, plays a role in evaluating the statistical performance of these indicators. Overall, this paper gives more clarity to the predictive performance of these
indicators as well as practical guidance on how to use them in a trading environment as generators of market entry ideas.

**Keywords:** technical analysis; backtesting; permutation test; financial markets; commodity futures; contract rollover.

## 1 INTRODUCTION

DeMark indicators are chart-based indicators that generate market entry signals. This family of indicators is, commercially speaking, quite popular. It is possible to make use of them as an upgrade in leading financial market terminals such as Bloomberg Professional and Thomson Reuters, which, combined, account for roughly 60% of the market share (Stafford 2015). Despite this, no previous study has analyzed the effectiveness of these indicators, although there are other studies available on simpler chart patterns (see, for example, Lo et al 2000). This paper gives more clarity to the predictive performance of these indicators by testing if they can time exceptional price moves and suggest the right side of a trade. Our aim is to understand what we should expect from the price development just after a completed signal. This should help us to manage a trade right from the start, e.g., by delaying the market entry, or by holding to a temporary initial loss that is expected to pay back shortly.

The first step toward an observed price predictability is to compute the performance of each market entry signal for a sliding number of holding days. To complete the picture, the results have to be compared with the market performance. This is done via permutation tests: we create batches of random signals that carry the same features as the original signal and bypass several limitations of a simpler buy-and-hold comparison. For example, the number of market entries and holding days are the same as in the tested signal. If, statistically speaking, the performance of the original signal goes significantly beyond what can be expected from equivalent sets of random signals, then the indicator can time a sizable price move that is yet to come within a limited number of holding days. The tests carry three performance measures, but the results are all quite similar. Therefore, this paper focuses on one of them: the (mean) returns per trade. The most suitable indicator for a given market should jointly maximize, for example, the (mean) returns per trade and the percentage of holding days following entry signals that show statistically significant performance. In most cases, there is a limited range of holding days for which an indicator has predictive power. This might sound counterintuitive, but it means that significantly exceptional price moves are yet to come, as the entry signal suggests; however, they may come with a delay and only last for a few days.

DeMark indicators are a small corner of a vast topic called technical analysis (TA). They are tested on commodity futures contracts, the basic type of exchange-traded...
financial instrument for commodities that is right at the intersection of physical and financial trading. Academic studies of TA on commodity futures markets are certainly not something new; some are quite old (Lukac and Brorsen 1990; Lukac et al 1988; Roberts 2003), while others are more recent and rather optimistic about the technical trading content of commodity futures prices (Miffre and Rallis 2007; Szakmary et al 2010). Other modern approaches are more pessimistic (Chinn and Coibion 2014; Marshall et al 2008). We observe in our study that different conclusions can be due to different constructions of continuous price series. Once a specific contract is used to determine the price for a given trading day, there are multiple options regarding when (Carchano and Pardo 2009; Ma et al 1992) and how (Masteika and Rutkauskas 2012; Masteika et al 2012; NASDAQ 2013; Pelletier 2011) to roll over to the next contract. In particular, a roll on the expiry should be avoided because it may underestimate the predictive power of DeMark signals.

We continue with an introduction to TA and its use in commodity markets; this is followed by a contextualization of the backtesting of predictive market signals. In TA, prices have always been the primary reference of past market activity with which so-called technicians could attempt to predict future market sentiment. Recent books and literature surveys (Chan 2013; Irwin 2007; Menkhoff and Taylor 2007; Pardo 2008) tend to conclude that there is some value and predictive power in TA. This contrasts with the earlier, more skeptical perceptions of academic researchers based on various forms of the efficient market hypothesis (EMH; see Fama (1995)). One belief of TA is that, like physical objects, prices have inertia: when at rest, they often stay approximately at rest, and when in motion, they often stay in motion (Widner 1998). This is exemplified, for instance, in the probabilistic mechanical view of market movements proposed by Andersen et al (2000). Price levels and price ratio relationships between highs and lows give price-based forecasting techniques that try to identify conditions in which prices are in motion along the trend. For example, prices tend to bounce on support and resistance levels (Garzarelli et al 2014). However, the crossing of these levels is interpreted as prices moving in a way that is likely to continue along the trend (Kosar 1991). There is also another approach to predicting future market movements: time-based forecasting (Coles 2011; Miner 1991). This class of methods tries to identify patterns in time series that should repeat over time.

Multiple factors contribute to the deviation of commodity prices from fundamental values for periods long enough to disturb normal decision-making processes (Filimonov et al 2014; UNCTAD 2011). This statement is supported by recent empirical evidence on the existence of speculative bubbles in several commodity markets (Gilbert 2010; Phillips and Yu 2010; Sornette and Cauwels 2014, 2015a; Sornette et al 2009). In this context, the use of TA can be justified as a tool for price discovery. In fact, TA has always had a significant and consistent user cohort among commodity traders (Billingsley and Chance 1996; Lukac et al 1988; Menkhoff and
Taylor 2007; Smidt 1964). As we have said, there are several factors that could push the price away from its fundamental value. For example, market participants cannot always enter positions when a price correction is yet to come (Gromb and Vayanos 2010) because market exposures in trading books are limited by capital constraints (Shleifer and Vishny 1997) and internal risk limits. In addition, even well-informed commodity traders must formulate price expectations based on partial or uncertain data (Gorton et al 2007; Khan 2009); this stimulates the use of rational herding behaviors, which have been described by Devenow and Welch (1996), Bikhchandani and Sharma (2001) and Hirshleifer and Teoh (2003). Herding behavior can also be irrational. Noise traders keep or adjust their positions independently of any changes in commodity fundamentals, and based on judgmental biases (Ariely 2010; Grinblatt and Han 2005; Penteado 2013); positive-feedback mechanisms (Sornette and Cauwels 2015b); simple TA rules, which can also be easily understood by traders with no fundamental understanding (Gehrig and Menkhoff 2006); and cross-asset strategies (Tang and Xiong 2010).

The first known attempt at backtesting trading signals using historical data was undertaken by a professional astrologer from Antwerp, who in 1540 tried to distinguish himself by testing his astrological system, which he said could foretell local commodity prices (Ehrenberg 1928; Lo and Hasanhodzic 2010). His main idea was to use stars as a way to generate random patterns. Centuries later, the statistical problem of insignificant predictive evidence in TA became well known (James 1968; Jensen 1967). Major methodological innovations arrived later still, in parallel with increased and more accessible computational power. In statistical testing, the replacement of predefined returns distributions with simulated distributions goes back to the bootstrap approach (Brock et al 1992). Data snooping used to be checked using out-of-sample data until White (2000) developed an in-sample “reality” check. All of these technical improvements make the study of TA’s predictive power more rigorous, but we should keep in mind that the determination of the predictability in financial markets is far from being 100% accurate (Zhou et al 2012). There is an economic upper bound to TA’s predictive power, which is based on the risk–return principle (Ross 2005; Zhou 2010).

The paper is organized as follows. Section 2 describes the DeMark indicators used to generate entry and trading signals. Section 3 goes into more detail regarding commodity futures, and explains the problem of rolling commodity futures contracts. In addition, it suggests a method for creating continuous daily returns starting from separate futures contracts. In Section 4, different ways of evaluating the performance of entry/trading signals are discussed, with a focus on Monte Carlo permutation tests. Using a two-dimensional framework based on our Monte Carlo permutation tests, the main results from DeMark backtests on rolled commodity futures are then presented. Section 5 concludes.
ALGORITHM 1 Entry signals.

1: for each (new) \( t \)-th price bar:
2: procedure DeMark\((P_c, P_h, P_l, n, m, q, p, k, \text{options})\)
3: update Setup's counter \( F \) eq. 2.1
4: if \( s = m \) then
5: compute Setup’s range & \( R_w \) eq. 2.2, 2.3
6: update Support/Resistance levels
7: end if
8: if indicator = ST then
9: if no open positions then
10: if \( P_c(t) > \text{res}(t) \) then
11: open new long position at \( t + 1 \)
12: else if \( P_c(t) < \text{sup}(t) \) then
13: open new short position at \( t + 1 \)
14: end if
15: end if
16: else if indicator = Sequential then
17: if \( s = m \) then
18: if no active Countdown phase then
19: activate new Countdown phase
20: reset Countdown’s counter \( (c = 0) \)
21: end if
22: end if
23: if active Countdown phase then
24: check recycle and ending conditions
25: if Combo then
26: update \( c \) eq. 2.7
27: else (normal Countdown)
28: update \( c \) eq. 2.5, 2.6
29: end if
30: if \( c = p \) then
31: open new position eq. 2.8
32: end if
33: end if
34: end if
35: end procedure

2 DEFINITION OF DEMARK INDICATORS TO BE TESTED

This is a family of indicators developed over time by Tom DeMark. It collects many revised versions of traditional price-based indicators, such as moving averages, trend lines, price ratios and Elliot waves, and also includes other indicators, among which “Sequential” is arguably the most renowned. When reading DeMark (1997) and Perl
(2008), one may feel that everything is described but nothing is explained. This further motivates us to explore the performance of these indicators, although we are aware that the rationale behind some choices (e.g., the values assigned to the parameters) will stay unexplained. The indicators Sequential, Combo (Sequential’s main variation) and Setup trend (ST) all generate long and short entry signals based on Algorithm 1. This pseudocode uses the following inputs: historical closing, high and low prices \((P_c, P_h, P_l)\), a set of parameters given by DeMark \((n, m, q, p, k)\) and the option to choose among slightly different versions. Our description focuses only on long entry signals; short entry signals can be derived by applying symmetrical conditions. See Appendix A (available online) for a summary of the parameters used in the backtests.

In row 3 of Algorithm 1, whenever a new price bar is available at time \(t\), a counter will be updated. Its value \(s\), which is initially set to zero, is increased by one \((s = s + 1)\) each time the following long condition is fulfilled:

\[
P_c(t) < P_c(t - n) \quad \text{for each (new) } t, \tag{2.1}
\]

with \(n = 4\), according to DeMark. The counter increases only if there are consecutive closing prices satisfying (2.1); otherwise, it is set back to zero. A parallel counter is running based on a symmetrical short condition. If the long counter increases, then the short counter must be reset to zero, and vice versa. DeMark sets \(m = 9\), and whenever \(s = m\), a Setup is complete. However, this will not reset the counter back to zero. A Setup ends only when the number of consecutive closes cannot be increased. Figure 1 shows an example of a long Setup.

Each time \(s = m\), the price bars related to the newly completed Setup are used in row 5 to determine the range of the Setup,

\[
\left[ \min_{s=1,\ldots,m} (P_l)_s; \max_{s=1,\ldots,m} (P_h)_s \right], \tag{2.2}
\]

and the width of the range,

\[
R_w := \max_{s=1,\ldots,m} (P_h)_s - \min_{s=1,\ldots,m} (P_l)_s. \tag{2.3}
\]

The same price bars also update support (sup) and resistance (res) levels:

\[
\text{res}(t) := \max_{s=1,\ldots,m} (P_h)_s, \tag{2.4a}
\]

\[
\text{sup}(t) := \min_{s=1,\ldots,m} (P_l)_s, \tag{2.4b}
\]

for each completed Setup. New long Setups update resistance levels (see Figure 2), whereas short Setups update support levels.

Rows 8–15 of Algorithm 1 describe long and short entry strategies for ST. For example, a new long position will be entered on the next traded day \((t + 1)\), at \(P_c(t + 1)\),
as soon as the latest closing price breaks its resistance level \( P_c(t) > \text{res}(t) \), as in Figure 3.

Starting from row 16, the pseudocode explains how Sequential can generate entry signals. At each new completed Setup \( (s = m) \), a new Countdown phase starts, unless one is already active (rows 17–22). Like the Setup phase, the Countdown phase can be long or short and has its counter set to zero as soon as a new Countdown begins \( (c = 0) \). For each new \( t \)th price bar, counter \( c \) will increase by one on a long Countdown only if the following conditions are met (rows 27–28). Given \( t \),

\[
\begin{align*}
\text{(standard)} & \quad P_c(t) \leq P_l(t-u), \\
\text{(aggressive)} & \quad P_l(t) \leq P_l(t-u),
\end{align*}
\]

and \( u = 2 \), as suggested by DeMark. Only one of the two variants in (2.5) should be used. Unfortunately, as already mentioned in the abstract, signals are sparse; we opt for the aggressive version in our backtests because it maximizes the number of entry signals. These conditions are very similar to those in (2.1): \( P_l \) replaces \( P_c \), \( u \) replaces \( n \) and, here, the equality condition is also accepted. In addition, for both cases, the \( p \)th bar completes the Countdown if, given the \( k \)th price bar,

\[
P_l(p) \leq P_c(k).
\]

DeMark suggests setting \( p = 13 \) and \( k = 8 \).
If (2.6) is not verified immediately, then the completion of this phase is postponed until this condition is met in one of the later bars. Unlike in the Setup, what matters in the completion of the Countdown is the total number of bars $p$ fulfilling (2.5), and not the consecutive number (see Figure 4). Counter $c$ increases based on conditions that are independent of the Setup, although, during the Countdown (or its alternative, the Combo), there are additional checks based on the Setup that can restart (“recycle”) the phase or, in a worst-case scenario, end it before its completion (ie, row 24).

**New opposite Setups.** A completed Setup in the opposite direction will restart the existing Countdown. For example, a short Setup is completed while a long Countdown is still building up. The long Countdown will end immediately, to be replaced by a short Countdown with $c = 0$.

**Crossed support and resistance levels.** For example, in case of a buy, if $P_c(t) > \text{res}(t)$, then the Countdown is stopped and canceled. This condition is similar to that which generates entry signals using ST, but, here, closing prices are replaced by high and low prices.

**New Setups.** If the new Setup has a $P_c$ within the range of the old Setup, then the current Countdown is kept going. Otherwise, if there is a new completed Setup in the same direction as the old one, the Countdown is reset ($c = 0$) only if $R_{\text{new}} > R_{\text{old}}$. 
FIGURE 3  In June 2013, WTI’s $P_c$ breaks its resistance level ($P_c(t) > \text{res}(t)$). Therefore, ST generates a new long entry signal.

![Graph showing price movements and support resistance levels.](image)

FIGURE 4  A completed Countdown on light crude oil.

![Graph showing price movements and countdown setup.](image)
Instead of using (2.5) and (2.6), it is possible to increase counter $c$ by one if all of the following Combo conditions are fulfilled (rows 25–26). Given $t$,

\begin{align}
P_c(t) &\leq P_l(t - u) \quad (= \text{Equation (2.5)}), \\
P_l(t) &\leq P_l(t - 1), \\
P_c(t) &< P_c \quad (\text{of the previous Combo bar}), \\
P_c(t) &< P_c(t - 1).
\end{align}

(2.7a) \hspace{1cm} (2.7b) \hspace{1cm} (2.7c) \hspace{1cm} (2.7d)

A major difference with the normal Countdown is that the bar check in (2.7) starts from the first bar of the Setup instead of the last (see Figure 5).

Once the Countdown (or Combo) is complete at time $t$, an entry strategy determines when to start a long entry signal (row 31). The aggressive strategy enters a long position immediately on the next traded day ($t + 1$) at $P_c(t + 1)$. However, the conservative strategy enters at $P_c(t + 1)$ as soon as

\[ P_c(t) > P_c(t - n). \]

(2.8)

Equation (2.8) is similar to the Setup check (2.1), but it has a reversed test direction. We chose the conservative strategy in our backtests because it is the default configuration and tries to optimize the timing by delaying an entry signal for a few price bars, if needed (Figure 5). Anyway, the choice of the entry strategy does not seem to be the
key factor for the performance of the indicator, because, at this point, an entry signal can only be delayed, not canceled.

Sequential (in its traditional Countdown or its alternative Combo version) is a time-based indicator that tries to identify areas of trend exhaustion that will lead to price reversals. It is made up of two Sequential phases. The first is the Setup, which tries to capture price momentum; this is followed by the Countdown, which looks for momentum exhaustion. ST is a price-based indicator that uses the Setup to determine support and resistance levels. When these are crossed, prices are in motion and should continue along the trend. The Setup phase (rows 3–7) is a common starting point for both Sequential and ST entry signals.

To generate a long entry signal with Sequential, the Setup must first identify a bearish momentum in the market. To do so, closing prices (or settlement prices for derivatives) are compared with the close $n$ bars earlier (2.1). The idea of using $n$-days momentum to avoid noise is not unique and can also be found in Chan and Lin (2004). A long Setup is completed when there are $m$ consecutive closes, with each one less than the corresponding close $n$ bars earlier. Here, the goal is to identify a continuous negative price velocity, ie, a negative trend. There can still be price oscillations, but an $n$-days momentum guarantees that the amplitude is small enough and the period is short enough that there is no interruption of the Setup. According to DeMark, a negative price velocity is measured over four time periods ($n = 4$), and it has to be maintained for nine consecutive periods ($m = 9$) in order to identify a trend. We think that the choice of a four-days momentum is in line with the behavior of many traders, who are not necessarily interested in daily price moves as much as they are in price moves over a few trading days. Unsurprisingly, risk management reports built for traders are often sent out on a weekly basis in order to limit noisy information and to be in line with the traders’ way of thinking. If $m \leq 5$, then we are sure the closing price of the last Setup bar is lower than the first bar, which means a negative price velocity has actually moved the price down.

Setting $m$ to 9 means adding additional trend checks to the first five bars; the closing price will trend down with continuity, and the price from the first to the last bar will have gone further down, because $P_c(9) < P_c(5)$ while $P_c(5) < P_c(1)$. According to this interpretation, $m$ depends on $n$, while the latter is given a value in line with the traders’ way of looking at price moves; all of this makes the Setup a very general method to identify incipient trends. Larger oscillations are tolerated during the Countdown because the market price is still trending (with a negative price velocity), but it is decelerating. Deceleration is by no means constant, and this translates into an alternation of negative and positive price velocities. Equation (2.5) deterministically suggests that, after $p = 13$ negative velocities, the trend is finished and ready to revert as soon as the current price level of bar $p$ is below or approximately the same as it was on bar $k = 8$ (2.6). Therefore, prices can continue to be bearish, or
they can go sideways, while the trend exhaustion pattern is building up. Since larger oscillations are tolerated, price velocities have to be measured over a shorter time frame to capture velocity oscillations. This may be the reason why the Countdown and Combo’s $u = 2$ is smaller than the Setup’s $n = 4$. However, the fixed number of $p$ negative moves before prices start to recover may be dependent on specific markets within a defined asset class. It is reasonable to think that DeMark did not focus his studies on commodity markets, but rather on stock markets, given the systematic use of closing prices instead of settlement prices, which is a common term for commodity futures and swaps.

Note that long ST signals are based on bearish momenta. There is an acceleration from a zero price velocity to a sustained negative velocity each time a long Setup is complete. Despite a market force pulling the price down, if the current price has the strength to push itself above the resistance level (the highest price level within the latest completed long Setup), then the price is supposed to have enough inertia to continue its motion along the upward trend. When there is an ongoing long open position, we could also use a long completed Setup (ie, a bearish market force that sets a new resistance level) to exit the position. This is visible in Figure 3: when the price returns into the shaded area, the current long position should be closed. Unfortunately, there is no exit signal for Sequential; therefore, we will only backtest our indicators on their entry signals.

Last, a symmetrical algorithm generates a short entry signal for both Sequential and ST. In reality, there is no symmetry between uptrends and downtrends in the markets (Benyamini 2009). In fact, selling pressure can take long pauses, but when prices drop they tend to do so at a high and constant velocity. However, buying pressure is relatively even and generates slower and longer uptrends. In commodities, bottoms have larger price oscillations than tops. Given these differences, Sequential seems more suited to long entry positions just after fast price drops have occurred. In the case of long-lasting, self-sustaining uptrends, the risk for Sequential is to repeatedly suggest short entry positions while the trend is still ongoing.

3 COMMODITY FUTURES DATA

3.1 Futures data

We test DeMark indicators over twenty-one commodity futures markets and ten years of data (January 2004–January 2014). Some idiosyncratic properties of these markets can be found in Appendix B (available online). In this paper, we limit our analysis to outright positions. The performance is computed on daily settlement prices, on which we calculate daily returns, while the entry signals are generated using daily price bars that contain intraday information such as daily opening, high and low prices. Complete
TABLE 1  List of commodity futures used for backtesting.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Exchange</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Wheat</td>
<td>Grain</td>
<td>CBOT</td>
</tr>
<tr>
<td>2 Corn</td>
<td>Grain</td>
<td>CBOT</td>
</tr>
<tr>
<td>3 Oats</td>
<td>Grain</td>
<td>CBOT</td>
</tr>
<tr>
<td>4 Soybean meal</td>
<td>Grain</td>
<td>DM</td>
</tr>
<tr>
<td>5 Cocoa</td>
<td>Soft</td>
<td>ICE US</td>
</tr>
<tr>
<td>6 Coffee C</td>
<td>Soft</td>
<td>ICE US</td>
</tr>
<tr>
<td>7 Sugar 11</td>
<td>Soft</td>
<td>ICE US</td>
</tr>
<tr>
<td>8 Cotton 2</td>
<td>Soft</td>
<td>ICE US</td>
</tr>
<tr>
<td>9 Light crude</td>
<td>Energy</td>
<td>NYMEX</td>
</tr>
<tr>
<td>10 Natural gas</td>
<td>Energy</td>
<td>NYMEX</td>
</tr>
<tr>
<td>11 Heating oil</td>
<td>Energy</td>
<td>NYMEX</td>
</tr>
<tr>
<td>12 Brent crude</td>
<td>Energy</td>
<td>ICE EU</td>
</tr>
<tr>
<td>13 Natural gas</td>
<td>Energy</td>
<td>ICE EU</td>
</tr>
<tr>
<td>14 Gasoil</td>
<td>Energy</td>
<td>ICE EU</td>
</tr>
<tr>
<td>15 Aluminium</td>
<td>Industrial metal</td>
<td>SHFE</td>
</tr>
<tr>
<td>16 Copper</td>
<td>Industrial metal</td>
<td>COMEX</td>
</tr>
<tr>
<td>17 Copper</td>
<td>Industrial metal</td>
<td>SHFE</td>
</tr>
<tr>
<td>18 Gold</td>
<td>Precious metal</td>
<td>COMEX</td>
</tr>
<tr>
<td>19 Silver</td>
<td>Precious metal</td>
<td>COMEX</td>
</tr>
<tr>
<td>20 Platinum</td>
<td>Precious metal</td>
<td>NYMEX</td>
</tr>
<tr>
<td>21 Palladium</td>
<td>Precious metal</td>
<td>NYMEX</td>
</tr>
</tbody>
</table>

daily price bars were not available for London Metal Exchange contracts. Some other contracts had to be excluded as well; for example, CBOT soybean oil, soybean meal and NYMEX gasoline US could not cover the whole tested period. Table 1 lists all futures contracts chosen for backtesting. There is not much flexibility in the choice of the data period, because we want as many years as possible to maximize the number of entry signals, which are sparse (on average, between two and five per year); however, it is not possible to include older data without reducing the number of commodities. Meanwhile, if a new price history builds up, we should set it aside for out-of-sample testing. Anyway, the data period is still quite heterogeneous and includes uptrends and downtrends before and after the great recession in 2007–8, which is our structural break. A detailed analysis of each commodity price would be informative for the reader, but it is not essential for performance testing, because the methodology described later self-corrects the results based on the uptrends and downtrends in the markets.

This paper essentially deals with daily data. Commodity futures can be traded intraday, but this is as far as we can get if we want to generalize this study to most
commodity markets, which are traded over-the-counter (OTC) and do not have sufficient liquidity (compared with all other asset classes) to be part of a high-frequency setting. For most products, it is difficult to enter and exit positions on an intraday basis, and even getting daily highs and lows from a broker can be very challenging. Front energy futures contracts traded on NYMEX or ICE, like those in this study, have an open interest (OI) of a few hundred thousand contracts, while the most liquid OTC energy products have an open interest between 1000 and 10,000 front-month contracts. Even if the focus is just on futures, the importance of high-frequency data in commodities would still be debatable because of the role played by physical hedgers. The overall impact of these players is clear only at the end of the day. In fact, physical markets are based on end-of-day assessments by pricing agencies such as Platts, Argus and Icis, and when physical deals are pricing, hedging orders are sent to the exchange to trade at the daily settlement price. During the day, the impact of hedgers generates noise: new exchange-for-physical deals (EFPs) or, perhaps, contractual triggers due to physical liftings, which have no direct relationship with the current flat price moves, force the hedgers to reposition immediately in the market. In addition, a sudden increase in expected forward oil production due to an expected increase in long-term demand (e.g., equivalent to ten Aframax cargoes or, equivalently, 10,000 crude futures contracts) could immediately generate strong selling pressure on crude futures due to the need to hedge by immediately shorting the futures.

3.2 Rolling futures contracts

A continuous price series of traded futures must have one contract selected for each trading day. Prices belong to the front part of the curve \((M, M+1, M+2)\) to capture spot price movements. This part of the curve is also the backbone of physical trading because it is used to evaluate and hedge unsold material. It is not possible to stick to one contract for the whole backtested period, because futures contracts can only be traded for a few years and cannot cover the whole backtest. When a contract starts being traded, it represents the most long-term future price expectation with a weak dependency on front price changes, a lower volatility compared with the front contract and a lower liquidity. The simplest way to build a continuous flat price curve is to roll following the last traded day of the latest expired contract. A more sophisticated method should anticipate the roll, because contracts show abnormal volatility in the final weeks of their lives (Samuelson 1965). Carchano and Pardo (2009) and Ma et al (1992) suggest rolling the front contract one to two weeks before maturity or on the first day of the expiring month. As an alternative, they also recommend staying on the most liquid contract, for example, by rolling from \(M+n\) to \(M+n+1\) as soon as the open interest on \(M+n+1\) is higher than on \(M+n\). Data providers suggest similar solutions (Thomson Reuters 2010), with the addition of rolling methods that
The black (unadjusted) curve is a generic continuous flat price curve, generated by one of the rolling strategies described in the text. Given the price level and the market’s contango, it could also represent a very steep front contango market on ICE gasoil after a flat price crash. Trading day 0 is the most recent trading day of the time period, and the plot goes backward in time, from right to left. The black curve shows price discontinuities on the rolling day. The blue curve, which is obtained, as explained in the text, to be neutral to contract jumps, is used to generate signals and test their performance.

roll from $M$ to $M + 1$ based on weighted values. In this paper, the following rolling strategies have been considered:

1. from $M$ to $M + 1$, following $M$’s expiry;
2. from $M$ to $M + 1$, ten days before $M$’s expiry;
3. from $M$ to $M + 1$, when $\text{OI}(M) < \text{OI}(M + 1)$; and
4. always on the contract with maximum OI.

No strategy is permitted to roll back to the previous contract, ie, from $M + 1$ to $M$.

The results presented in this paper are based on rolling 3. While rolling 3 seems like a reasonable choice for our backtests (see Section 4.2, where we study the impact of the different rolling methods on the tests), there is still room to further investigate the impact of rolling strategies.

There is an additional issue that we have to address before presenting the extensive results of our tests. When each trading day has a contract linked to it, then continuous settlement prices still show misleading price discontinuities on the rolling day (see the black curve in Figure 6). DeMark indicators use daily flat price movements to generate entry signals; a price movement in the wrong direction could stop a signal before its completion. Therefore, any DeMark indicator should not consider price discontinuities coming from contract rolls as price moves that can generate profits.

A simple way to remove discontinuities would be to use local adjustments by weighting price values across multiple trading days. Nevertheless, there are more
computationally intensive adjustment techniques that fit better with our backtests. The black unadjusted curve in Figure 6 shows a steep front contango market; in fact, the front contract $M$ is rolled twice to a higher-valued $M + 1$ contract. Assuming the black curve is related to a long position that we want to keep, we should sell a lower-valued contract and buy the next contract that always trades at a higher price ($+13$ and $+15$). From a mark-to-market perspective, changing the contract creates no profit and loss (P&L) if we exclude transaction costs. For this reason, entry signals are being generated on the blue trading curve, which is neutral to discontinuities coming from the rolling process.

There are multiple ways to adjust discontinuities on continuous flat price time series (Masteika and Rutkauskas 2012; Masteika et al. 2012; NASDAQ 2013; Pelletier 2011). There is a general consensus that prices in the old contract should be adjusted to prices in the new contract. In this way, continuous prices that are later in time should be less affected by adjustments. However, this backward adjustment is more computationally intensive than a forward adjustment, because a change to a single rollover requires additional changes to all previous trading days.

Figure 6 uses backward adjustments and employs a discrete translation of the price by the adjustment $\Delta$ on the day of the roll such that, on day $d$,

$$\Delta := P_o(d) - P_c(d - 1). \quad (3.1)$$

The blue curve, which is neutral to rolling, is obtained by adding the adjustment $\Delta$ to all prices ($P_c$, $P_o$, $P_h$, $P_l$) prior to the rolling day.

Another method is the proportionally adjusted method. Instead of adding a fixed-quantity $\Delta$, the proportionally adjusted method multiplies all the prices prior to the rolling day by the quantity $\rho$:

$$\rho := \frac{P_o(d)}{P_c(d - 1)}. \quad (3.2)$$

Using such ratios has the advantage that negative prices are not possible by construction. However, they may create large reconstructed price fluctuations. This unwanted feature has led practitioners to favor the use of the $\Delta$ approach over the $\rho$ approach.

In this paper, continuous flat price curves have been backward adjusted using $\Delta$ quantities. This method generates negative prices on the trading series (in blue) for soybean meal, copper COMEX and copper SHFE. In other cases, adjusted time series are close to the null price. This is not a problem for trading time series because DeMark indicators only use relative prices to build up entry signals. However, this is a problem when evaluating the performance of strategies in comparison with that of the commodity, defined by its daily market returns at day $d$:

$$r_d := \frac{P_c(d) - P_c(d - 1)}{P_c(d - 1)}. \quad (3.3)$$
The blue curve in Figure 6 cannot be used directly to compute daily returns; this is because settlement prices \( (P_c) \) can be negative or close to zero values, and this would distort daily returns. We propose using the blue curve to compute the numerator in (3.3) while the denominator uses settlement prices from the black unadjusted continuous flat price curve. In fact, the blue curve shows correct changes in relative prices, while the black curve refers to the market’s absolute price levels.

Finally, cash management plays a role in the P&L: rolling to a contract with a higher value means that initial margin requirements will be higher, whether the position is long or short. Without considering any portfolio effect, higher collateral may force trading businesses to borrow more money, for which additional interest rates should be paid. In this paper, performance is measured on the blue trading curve (ie, daily market returns), while transaction and financing costs are not included.

4 MARKET PERFORMANCE OF DEMARK INDICATORS

4.1 Methodology

A positive trade record requires profit generation, but trades should also beat the markets on a risk-adjusted basis in order to be attractive to investors. If indicators generate signals that outperform the market, then those indicators are informative; in other words, they have predictive power.

In this paper, we study the predictive power of each separate indicator. This is not sufficient to guarantee the profitability of an indicator, but it is already a first step toward it. Profitability requires a much broader discussion than we can undertake here. In fact, traders use their intuition and research skills to enter new positions (eg, by combining several trusted fundamental and technical indicators), but profitable traders are also excellent exposure managers (Covel 2009; Original Turtles 2003). They handle multiple diversified strategies while always keeping their portfolio volatility under control, and they know the best moment to cut losses and lock-in profits. From this perspective, the role of a predictive entry signal is to give an edge to the trader.

One of the simplest ways to test the predictive power of entry signals is to compare their total cumulated performance, eg, the total cumulated returns or net asset value (NAV) versus a buy-and-hold strategy over the whole data period. Nevertheless, this approach might not be the best choice if we are willing to accept that buy-and-hold is a reference strategy in commodity trading. A general limitation of this setup is that each trade carries a different allocation of money depending on the results of the previous trades (Siligardos 2014), and this distorts the measurement of predictive power. In addition, the results would be very sensitive to the data period, especially since the signals we are testing are sparse and mostly out of the market. We would be comparing different things, and further investigations would be needed. We would
have to convince the reader about the choice of the data period, perhaps through a robustness analysis that describes the general statistical properties of each commodity and the related structural breaks. For example, we would have to comment on the uptrends and downtrends before and after the great recession in 2007–8 and how these might affect the results. The analysis should also be strengthened by out-of-sample and in-sample data-snooping tests (see Diebold 2015; Kuang et al 2014). To some extent, by comparing the distribution of (daily) returns conditional on entry signals to their unconditional counterpart, one could attempt to derive conclusions regarding the aforementioned comparison between trading strategies and the buy-and-hold (Lo et al 2000). The conditional versus unconditional returns approach weights every entry signal equally, is easy to implement and offers graphical intuition, although it cannot combine longs and shorts on the same graph. For example, Figure 7 suggests that long Sequential entry signals might overperform the cocoa futures market when the holding period reaches thirteen or fourteen days.

Unfortunately, with a $p$-value of 0.64 in the Kolmogorov–Smirnov test, the conditional distribution does not seem significantly different from the unconditional distribution when the number of holding days is set to fourteen. A limit of this test is that it only compares the maximum gap between the two cumulative distributions, without considering that each conditional quantile is overperforming its corresponding market quantile. In addition, the test assumes independent and identically distributed (iid) returns, which is not plausible for financial data (Lo et al 2000). So, the failure of the Kolmogorov–Smirnov test to reject the null hypothesis, which, in this case, suggests
that the long Sequential entry signals do not overperform the cocoa futures market, may just be due to its lack of power.

Unsurprisingly, the Monte Carlo permutation tests (Aronson 2007; Masters 2010) in Figures 8–10 provide a conclusion that is more in line with the previous graphical intuition. We will explain this later. For now, we will stick to this methodology, because, by design, it can handle sparse entry signals that are also held for short holding periods. In addition, the results are not biased by the trends and the structural breaks of the underlying commodities, and this makes the choice of the data period less critical. Finally, this testing keeps the returns in the same order as they appeared historically.

According to the null hypothesis $H_0$, the long, short and neutral positions of the signal are paired randomly with daily market returns. The alternative hypothesis $H_A$ supports the idea that the current pairing improves performance beyond what could be expected from randomness, which also means the indicator is intelligent and has predictive power over the market. The value of such random strategies that have the same characteristics as the prediction system to test, except for the timing, is also explained by Daniel et al (2009) in the context of avoiding selection biases, survival biases and look-ahead biases during backtests.

Randomized signals have several constraints. All possible permutations, including the candidate signal, must have an equal chance of appearing in real life and in the randomization process. For example, the total number of trades, longs, shorts and the total number of trading days must always be the same. An additional constraint in our study is that each trade must be held for the same fixed number of days, because we examine the predictive power of entry signals by sweeping the number of holding days. This means that, by design, the candidate signal contains trades with a fixed number of holding days. Moreover, to avoid possible interactions, trades should not overlap in time. Finally, each daily return needs to be paired with a long, short or neutral market position.

Compared with the bootstrap method, permutation tests have more requirements to fulfill. However, these tests overcome many bootstrap weaknesses. There is no assumption on the null hypothesis distribution used for the test, while in bootstrap tests the empirical distribution of the obtained sample is assumed to be representative of the whole population. Further, bootstrap tests re-sample returns with replacement while the trading signal is kept unchanged. This is how the null hypothesis distribution is generated. However, during permutation tests, daily returns are kept on their historical positions while the trading signal is permuted (without replacement). In this way, the intelligent part of the returns is left untouched together with its behavioral and statistical dependencies (e.g., autocorrelations).

The total number of signal permutations grows very quickly, given the number of trading days, entered trades and holding days for each trade. Let us assume that there
are a total of thirty trading days, and that the signal can only be long or out of the market. If, with all the mentioned constraints, there is only one trade that lasts three holding days, then there are only twenty-eight possible permutations. If, instead, there are five trades with the same duration, then the number of signal permutations goes quickly up to $\sim 10^{4.2}$. In this paper, 400 randomized trading signals are simulated from each candidate signal, on each test. It is a trade-off between quantile smoothness, size of confidence intervals and computational power. Permuted distributions are approximated; therefore, observed quantiles need to be adjusted according to confidence intervals before being used for hypothesis testing. Let $q$ and $\hat{q}$ be the true and the observed quantile, respectively. These transformations use a normally approximated binomial method in accordance with Conover (1999). A limitation of this method is that it requires large samples, but this is not a problem for our permutation tests. Its strength is that it can be applied to any quantile. Given the number of randomized signals $n_0$, the observed quantile $\hat{q}$, the confidence level $\alpha$ and $Z_\alpha$ as the $Z$-statistic (eg, $Z_{1-\alpha} \sim 1.65$ when $\alpha = 95\%$), the true quantile $q$ has the confidence interval $\hat{q} - \varepsilon \leq q \leq \hat{q} + \varepsilon$, with

$$
\varepsilon = \frac{Z_{1-\alpha} \sqrt{n_0 \hat{q}(1 - \hat{q})}}{n_0},
$$

(4.1)

where $n_0 = 400$, $\alpha = 95\%$ and $\hat{q} = 95\%$, $93.2\% \leq q \leq 96.8\%$. When a $p$-value refers to a theoretical $q = 95\%$ quantile, we will conservatively pick a $\hat{q} = 97\%$ measured quantile from the empirical distribution. To avoid approximation, using the “exact” Clopper–Pearson (1934) confidence interval is generally recommended; this is well described by Agresti and Coull (1998). Both methods provide the same rounded results for $\varepsilon$. All the quantile adjustments adopted in the permutation tests are listed in Table 2.

Each trading signal uses an aggregate measure to determine its performance. In our case, this measure can be a mean value over single positions, eg, a mean return over all the single trade returns, which we will denote $\text{Profit}_\text{trade}$. A further step in the analysis could be to substitute returns with a return-to-risk ratio, eg, the Risk-Return-Ratio (RRR) (Johnsson 2010), which is defined for a single trade as

$$
\text{RRR} := \frac{r}{\text{DD}_{\text{max}}},
$$

(4.2)

with $r$ and $\text{DD}_{\text{max}}$ being the (conditional) return and the maximum drawdown of the trade, respectively. $\text{DD}_{\text{max}}$ is defined as the largest (compounded, but it can also be found uncompounded) cumulative return within a defined time period. Drawdown-based measures are widely used in practice (Chekhlov et al 2005; Eling and Schuermann 2007). The overall RRR value will be computed as the mean value of the RRR values of each trade; we will refer to this as $\text{RRR}_{\text{trade}}$. The set of performance metrics
### TABLE 2  Quantile transformations in the Monte Carlo permutation test.

<table>
<thead>
<tr>
<th>Left tail</th>
<th></th>
<th>Right tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q = 2.5%$</td>
<td>$\hat{q} = 1%$</td>
<td>$q = 90%$</td>
</tr>
<tr>
<td>$q = 5%$</td>
<td>$\hat{q} = 3%$</td>
<td>$q = 95%$</td>
</tr>
<tr>
<td>$q = 10%$</td>
<td>$\hat{q} = 7%$</td>
<td>$q = 97.5%$</td>
</tr>
</tbody>
</table>

The null hypothesis distributions are approximated; therefore, observed quantiles need to be conservatively adjusted according to confidence intervals. $q$ and $\hat{q}$ are the true and observed quantiles, respectively.

is completed by a measure that can be computed only over multiple positions, such as the profit factor. This is a profit-to-loss ratio, which is defined for a basket of trades as

$$P_f := \frac{\sum \text{Profits}}{\sum \text{Losses}}. \quad (4.3)$$

Assuming that all trades have the same allocation of money, $P_f$ can be rewritten as the sum of all returns from winning trades divided by the sum of returns from losing trades:

$$P_f = \frac{\sum r^+}{\sum r^-} = \frac{N^+ \bar{r}^+}{N^- \bar{r}^-}, \quad (4.4)$$

where $N^+$ and $\bar{r}^+$ are the number of winning trades and their mean return, respectively. On the other hand, $N^-$ and $\bar{r}^-$ are the number of losing trades and their mean return. A trade is considered a winner when its gross return is strictly positive. Next, given the following definitions of payoff ratio and win ratio,

$$P_r := \frac{\bar{r}^+}{\bar{r}^-} \quad \text{and} \quad W := \frac{N^+}{N^+ + N^-}, \quad (4.5)$$

the profit factor can be written to obtain

$$P_f = \frac{W}{1 - W} P_r. \quad (4.6)$$

A trading signal generates profits when $P_f > 1$, and, by definition, the breakeven is reached when $P_f = 1$. The desired condition is often $P_f > 2$ (Harris 2009). Equation (4.6) highlights how signals with low $P_r$ must have a high $W$ to generate profits, as is the case for intraday trading. On the other hand, high $P_r$ values coupled with a
low win ratio (i.e., $W < 50\%$) can still be profitable. $\text{Profit}_{\text{trade}}$, $P_f$ and $\text{RRR}_{\text{trade}}$ are computed for each candidate and randomized signal. Performance measures such as $\text{Profit}_{\text{trade}}$ and $P_f$ use double-sided tests. When the observed profit factor is below the breakeven ($P_f = 1$) and the random distribution is performing significantly better, this condition is informative and may indicate that the direction of the trades should be flipped. $\text{RRR}_{\text{trade}}$ uses a single-sided test instead. $\text{RRR}$’s definition uses the maximum drawdown at the denominator. As a consequence, it is not possible to assign a symmetrical interpretation to this measure when returns per trade are, for example, all negative.

The results for long Sequential tests on cocoa are shown in Figures 8–10. In the current backtesting, a $p$-value of 10% translates into a possibly significant rejection of $H_0$; a value of 5% translates into a statistically significant rejection. The lightly shaded area represents possible statistical significance at the 90% level, and the darkly shaded area represents statistical significance at the 95% level. Both use conservatively adjusted quantiles by taking into consideration confidence intervals. Turning our attention to possible significance, the unadjusted quantiles on single-sided tests ($q = 90\%$) and double-sided tests ($q_1 = 5\%, \; q_2 = 95\%$) have been adjusted to $\hat{q} = 93\%, \; \hat{q}_1 = 3\%$ and $\hat{q}_2 = 97\%$. This means that the candidate indicator (in Figures 8–10, and in Table 3) is never statistically significant, because the use of conservative quantile transformations on all performance measures is always decisive in
FIGURE 9  $P_f$ permutation tests for long Sequential on cocoa.

FIGURE 10  $RRR_{trade}$ permutation tests for long Sequential on cocoa.
TABLE 3  Measured quantiles for long Sequential on cocoa.

<table>
<thead>
<tr>
<th></th>
<th>Profit\text{trade}</th>
<th>P_f</th>
<th>RRR\text{trade}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thirteen holding days</td>
<td>95%</td>
<td>95%</td>
<td>87%</td>
</tr>
<tr>
<td>Fourteen holding days</td>
<td>95%</td>
<td>92%</td>
<td>90%</td>
</tr>
</tbody>
</table>

FIGURE 11  Impact of rolling strategies on the indicator's Profit\text{trade} performance and statistical significance. [Figure continues on next three pages.]

4.2 Impact of the rolling method on performance results

Figure 11 examines the impact of different rolling strategies for long ST on soybean meal, short Sequential on cocoa, and long and short Sequential on natural gas ICE. When the roll is done on the expiry day, the number of statistically significant holding days decreases compared with rolling 3, and in two examples there is no statistical significance left. In practice, the rolling strategy plays a key role in evaluating the
statistical performance of a trading indicator similar to DeMark on commodity futures markets. Rolling 1 is more conservative: it tends to accept (rigorously speaking, it tends not to reject) the null hypothesis more easily than the other rolling strategies. This also means that a basic backtesting on commodity prices series using rolled prices based on rolling 1 may underestimate the predictive power of an indicator. Further, rolling 1 always has the lowest Profit\text{trade} values (in magnitude) among the four rolling strategies. Rollings 2 and 3 have a similar number of trades and a similar Profit\text{trade}, although the stability of predictive power across the number of holding days may vary. Rolling 4 always picks the most liquid contract, but the impact strongly depends on the futures market. Rolling 3 on cocoa is very similar to rolling 4. The reason for this is that the front contract is the most liquid until twenty to thirty days before its expiry; therefore, the two strategies roll very close in time. In contrast, rollings 3 and 4 show very different behaviors for natural gas ICE. This can be explained by the strong seasonality of natural gas contracts. The roll is discontinuous: some contracts are more important and can be used for long periods in the continuous flat price curve, while others can be completely ignored.
4.3 Synthesis of the performance tests

Three DeMark indicators have been tested to generate long and short outright entry signals on commodity futures. The number of holding days is swept, which allows us to study trade performances during the days that follow the entry signals, as we have seen in Figures 8–10.

Given a fixed number of holding days, as before, we consider the following performance metrics introduced in Section 4.1:

(i) mean return \( \text{Profit}_{\text{trade}} \);

(ii) profit factor \( P_f \), defined in (4.3); and

(iii) Risk-Return-Ratio \( RRR_{\text{trade}} \).

If such a performance metric is measured to go significantly beyond what can be expected from randomness by having a value within the shaded area of previous figures, then the indicator is predictive. Ideally, measured performance should be
within the darkly shaded quantiles for each holding day. In this way, positions can be entered with a delay while still being on the correct side of the trade. They can also be closed whenever the trader feels more comfortable.

However, in most cases, there is a limited range of holding days for which the indicator has predictive power. This may sound counterintuitive, but it means that a significantly exceptional price move is yet to come, as the entry signal suggests; however, it may come with a delay and only last for a few days. If decision makers are aware of these signal–market behaviors, they can have a better idea of how to manage the initial phase of a trade, eg, by delaying the market entry or by holding to a temporary loss that is expected to pay back shortly.

By design, the Monte Carlo permutation tests described in Section 4 use performance metrics to test the indicator’s predictive power over the holding days. To quantify the information contained in, for instance, Figures 8–10, we introduce the stability $\sigma$ of an indicator’s predictive power over a specific market, which is defined as the percentage of holding days following entry signals that show a statistically
TABLE 4  Across all commodities, each indicator enters the markets with the following yearly frequencies (ie, the average number of positions per year).

<table>
<thead>
<tr>
<th>Positions/year</th>
<th>Sequential</th>
<th>Combo</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>3.0</td>
<td>1.2</td>
<td>3.4</td>
</tr>
<tr>
<td>Mean</td>
<td>3.9</td>
<td>1.7</td>
<td>4.5</td>
</tr>
<tr>
<td>Maximum</td>
<td>4.6</td>
<td>2.2</td>
<td>6.1</td>
</tr>
</tbody>
</table>

significant performance, with 100% representing the optimal case. The most-suited indicator for a given market should jointly maximize $\sigma$ (this is the priority) and the performance measure, which we also call profit potential. The latter is computed as the average of $\text{Profit}_{\text{trade}}$ values that belong to statistically significant holding days, if there are any. This seems to be a simple but effective way of limiting the data-mining effect that would occur if the maximum $\text{Profit}_{\text{trade}}$ value within the fourteen holding days was chosen to represent the profit potential of an indicator.

This two-dimensional evaluation framework with $\sigma$ (ie, percentage of significant holding days) on the horizontal axis and profit potential on the vertical axis (ie, average $\text{Profit}_{\text{trade}}$) has been applied to the DeMark indicators over the twenty-one commodity futures markets listed in Table 1. The indicators are evaluated only for their long entry signals in Figure 12(a), only for short entry signals in Figure 12(b) and for both longs and shorts in Figure 12(c). The most easily manageable signal–market combinations are represented in the shaded areas of Figure 12.

DeMark indicators have sparse entry signals. Across all commodities, each indicator enters the markets with the frequencies shown in Table 4.

ST provides the highest number of trades and is predictive for a few markets, especially those with long entry signals. In addition, it correctly captures the direction of the market, and this is always true when both long and short positions are possible. Sequential and its Combo variant generate less entry signals than ST. As explained in Section 2, Sequential and Combo always need a new Setup before completing an entry signal, while ST uses the Setup phase only to update support and resistance curves. If we also consider the fact that these two indicators need to fulfill additional conditions during the Countdown or Combo phase, then it (ideally) takes around one month before an entry signal can be completed on Sequential or Combo. Quite interestingly, Sequential has shown statistically significant performance for either long or short positions on all commodity futures apart from platinum. Unfortunately, just like its Combo version, this indicator does not seem to be able to forecast the correct market direction. This limits our trading strategies, but a forthcoming price move in an unknown direction can still be made profitable, eg, by using options-based strategies. Looking at Figure 12(a), a long Sequential on energy products is more of a
FIGURE 12 Evaluation framework that shows for each indicator its $\sigma$ on the horizontal axis (percentage of significant holding days) and the corresponding profit potential on the vertical axis (average Profit\textsubscript{trade}).

The commodity market is specified by a number that refers to Table 1. Blue represents Sequential (Sq.), green is for Combo (Cm.) and red is for ST. Combinations with no predictive power are not shown here.
trend detector that a turning point detector, and the market is more likely to continue its negative trend than shift to a new positive trend.

Therefore, Setup can identify negative trends, but markets seem to have a price inertia that is too big to be captured by the current Countdown parameters. The good news is that signal–market combinations, which in Figure 12 appear in a stable but unprofitable area, can still be informative if interpreted correctly.

So far, Figure 12 has provided a high-level overview of the behavior of the indicators on several markets. The framework can be studied in more detail by comparing results from different performance measures within the permutation test setup, as in Figure 13. Conditional distributions of returns versus unconditional returns can be used as an additional countercheck.

For example, Figure 12(a) highlights a good performance for long ST entry signals on soybean meal. A more in-depth analysis based on Figure 13 shows that Profittrades grows regularly over the holding period; at the same time, it stays steadily in the
shadowed area, which represents a possible or statistically significant result. Further, the quantiles of conditional returns are also steadily above the corresponding quantiles generated from unconditional market returns. These observations are partially confirmed by additional permutation tests using different metrics. While results are similar when $P_f$ is being used, $RRR_{trade}$ is in the region of significance, too, but less frequently. The same kind of thinking should be applied to the other examples. Sequential on cocoa is shown as a successful example for short positions. If we also recall the interesting results for long positions previously described in Section 4, then Sequential is also a predictive indicator when both long and short positions are considered; this is confirmed by Figure 12(c). The last example in Figure 13 exhibits statistical significance for long and short ST entry signals on wheat and shows how results may depend on the holding period. There is good predictive power across all metrics, but this only appears eleven or twelve days after the position has been entered.
5 CONCLUSION

Overall, this paper has taken a step away from conventional backtesting approaches. The most common way of studying the weak form of market efficiency is to test a wide range of technical indicators over many markets by using simple buy-and-hold-based metrics, and to legitimize the best results by using standard data-snooping tests, such as the Diebold–Mariano test. In our case, we tailor a general interdisciplinary tool (a permutation test) to the needs of a few specific indicators that generate sparse entry signals. We also analyze the impact of different commodity futures’ rolling strategies and performance metrics, with a focus on testing what happens immediately after a new entry signal is generated.

The specific indicators we study are DeMark indicators. These are intriguing because their layout has been described in multiple books (DeMark 1997; Perl 2008), and they are often used by practitioners to time the market, but the rationale for these indicators is not entirely clear, and the choice of their parameters is not really motivated and/or supported by data.
This paper accepts the DeMark indicators as they are and focuses on providing a better understanding of how they might perform as generators of market entry signals. More specifically, we set out to test if there is some predictive power in them: whether they can time exceptional price moves and suggest the right side of a trade. As decision makers, we would like advice on how to manage the initial phase of a trade, e.g., by delaying the market entry, or by holding to a temporary initial loss that is expected to pay back shortly. Profitability measurement is out of scope, since it also depends on exposure management techniques that lock-in profits and cut losses.

The first step is to describe the signal generation process of these indicators. In this paper, we study three specific indicators: Sequential, Combo and ST. DeMark indicators (especially Sequential) are not so easy to replicate because they involve multiple conditions that overlap in time. Therefore, we introduce a pseudocode to facilitate the understanding and reproducibility of these signals.

The testing is based on daily prices. Although commodity futures can be traded intraday, daily data is the most natural way to start looking at commodities. In fact, daily futures settlement prices are the best reflection of the effects of all market players, including the hedgers, who typically send orders to the exchange to trade at the daily settlement price. Further, most commodity markets are traded OTC, so they do not have sufficient liquidity (compared with all other asset classes) to be part of a high-frequency setting.

Our first conclusion is that these entry signals are sparse. New signals materialize mostly between one and five times per year. This observation affects the way in which the backtesting is designed. The data period has to be as long as possible so that we are able to compute some statistics on the entry signals. The starting date is January 2004 (giving us ten years of data); we prefer not to go further back because DeMark signals also need daily highs and lows, and having older price histories would mean...
that we would have to exclude other futures contracts from our study. Nevertheless, the data period is still quite heterogeneous and includes uptrends and downtrends before and after the great recession in 2007–8, which is our structural break.

A performance comparison with a buy-and-holder strategy over the entire data period is one of the simplest performance tests. The comparison can be done by calculating a metric like the NAV, or by looking at the distribution of (daily) returns conditional on entry signals versus their unconditional counterpart to test if average returns are increasing, if standard deviation or kurtosis are decreasing and if skewness is being pushed into a positive territory. If we compared strategies that are mostly out-of-the-market (few entry signals that are held for short holding periods) with a buy-and-hold that is always in-the-market, the analysis would fall short of convincing the reader about the performance of the indicators; data-snooping tests would become necessary. The comparisons would depend on the trends and the structural breaks of the underlying data periods and would be sensitive to the performance of a limited number of entry signals. On top of that, a strategy such as buy-and-hold, which is in-the-market for more days, would also carry more risk. We opt for permutation tests, which bypass several problems, including the ones just mentioned, and are suited to studying what happens immediately after the completion of these sparse entry signals over different trade holding horizons. In this setup, all possible permutations, including the candidate signals, must have an equal chance of appearing in real life and in the randomization process. Initially, the tests are based on three performance metrics. Mean return and profit factor always provide very similar results, while a risk-adjusted metric, such as the risk–return ratio, may lead to slightly different conclusions, but with no meaningful differences. For this reason, a high-level investigation is only shown for the mean return in Figure 12.

Multiple rolling strategies have been investigated; however, only one is used to test DeMark signals, and this rolls from the front contract to the next when its open interest is lower than that of the nearby contract. Other rolling strategies that slightly anticipate the expiry may be used as well, with similar results (see Figure 11). In particular, our paper suggests avoiding a roll on the expiry because it may underestimate the predictive power of these signals.

Our results show that, in most cases, there is a limited range of holding days for which the indicators have predictive power. This may sound counterintuitive, but it means that a significantly exceptional price move is yet to come, as the entry signal suggests; however, it may come with a delay and only last for a few days before becoming a nonexceptional move again. Armed with this information, a decision maker can better decide when and how to enter the markets. Sequential is an interesting case, because there are statistically significant price moves following either long or short entry signals on all twenty-one commodities save one. Although this signal
is designed to time trend reversals, in several cases the exceptional price move maintains the direction of the trend. This happens for expected trend reversals on energy futures: negative trends would rather continue after the signal instead of reverting back. These results contradict the design of the indicator and make it difficult to grasp the economic rationale behind it, which is assumed to be obvious. As a consequence, it becomes challenging to use Sequential, as it is designed to also offer some insight into the data-generating process for commodity futures prices and their times series properties. Nevertheless, we could still try to do so by exploring different values for the parameters, perhaps by sweeping Sequential parameters \( p, q \) and \( k \); although we would really have to pay attention to the data-snooping or selection bias. The higher the number of indicators being tested on the same historical data set, or the higher the number of data sets being tested on the same indicator, the higher the probability that luck will have an impact on the most statistically interesting results. In our tests, the results do not seem to be driven by luck due to a relatively low number of combinations (three trading rules over twenty-one data sets). In order to approach the problem gradually, without creating de facto new versions of the same indicator, we could first leave the parameters untouched and use the methodology in this paper to look at what happens before the completion of a new signal, to test if the signals are maybe arriving too late. The same methodology can also be extended to other entry and exposure management signals.

**DECLARATION OF INTEREST**

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

**ACKNOWLEDGEMENTS**

The ideas presented in this work originate from Marco Lissandrin’s Master’s Thesis, in fulfillment of his Master of Science degree at ETH Zürich. The full version can be viewed at http://bit.ly/2rl9Y5h.

**REFERENCES**


To subscribe to a Risk Journal visit subscriptions.risk.net/journals or email info@risk.net