Forecasting and Trading Commodity Contract Spreads with Gaussian Processes

Abstract

This paper examines the use of Gaussian Processes to forecast the evolution of futures contracts spreads arising on the commodities markets. Contrarily to most forecasting techniques which rely on modeling the short-term dynamics of a time series (e.g. ARIMA and most neural-network models), an appropriate representation of the input and target variables allows the Gaussian Process to forecast the complete future trajectory of the spread. Furthermore, as a customary outcome of using Gaussian Processes, the forecast includes not only the expectation of future spread prices (across time-steps), but their joint autocovariance matrix as well. We introduce a technique to exploit this joint autocovariance matrix in order to profitably trade spreads, based on maximizing an information ratio criterion between candidate entry-exit points and constantly monitoring the position with revised forecasts as the spread realization unfolds. This approach results in a qualitatively very different methodology than a classical mean-variance portfolio construction based on short-term forecasts, yielding models that do not overtrade yet react quickly to changes in market conditions. The method performs well out of sample, and we present extensive simultation results on historical data to validate the approach.

Key words: Bayesian methods, Finance, Financial markets, Nonparametric methods, Price forecasting, Time series

1. Introduction

Classical time-series forecasting models, such as ARMA models (Hamilton, 1994), assume that forecasting is performed at a fixed horizon, which is implicit in the model. An overlaying deterministic time trend may be fit to the data, but is generally of fixed and relatively simple functional form (e.g. linear, quadratic, or sinusoidal for periodic data). To forecast beyond the fixed horizon, it is necessary to iterate forecasts in a multi-step fashion. These models are good at representing the short-term dynamics of the time series, but degrade rapidly when longer-term forecasts must be made, usually quickly converging to the unconditional expectation of the process after removal of the deterministic time trend. This is a major issue in applications that require a forecast over a **complete future trajectory**, and not a single (or restricted) horizon. These models are also constrained to deal with regularly-sampled data, and make it difficult to condition the time trend on explanatory variables, especially when iteration of short-term forecasts has to be performed. To a large extent, the same problems are present with non-linear generalizations of such models, such as time-delay or recurrent neural networks (Bishop, 1995), which simply allow the short-term dynamics to become nonlinear but leave open the question of forecasting complete future trajectories.

Functional Data Analysis (FDA) (Ramsay and Silverman, 2005) has been proposed in the statistical literature as an answer to some of these concerns. The central idea is to consider a whole curve as an example (specified by a finite number of pairs (t, y_t)), which can be represented by coefficients in a non-parametric basis expansion such as splines. This implies learning about complete trajectories as a function of time, hence the "functional" des-

ignation. Since time is viewed as an independent variable, the approach can forecast at arbitrary horizons and handle irregularly-sampled data. Typically, FDA is used without explanatory time-dependent variables, which are important for the kind of applications we shall be considering. Furthermore, the question remains of how to integrate a progressively-revealed information set in order to make increasingly more precise forecasts of the same future trajectory. To incorporate conditioning information, we consider here the output of a prediction to be a whole forecasting curve (as a function of t).

This paper presents a solution to the problem of forecasting a complete future trajectory based on the use of *Gaussian Processes* (O'Hagan, 1978; Williams and Rasmussen, 1996; Rasmussen and Williams, 2006). They constitute a general and flexible class of models for nonlinear regression and classification. Over the past decade, they have received wide attention in the machine learning community, having originally been introduced in geostatistics, where they are known under the name "Kriging" (Matheron, 1973; Cressie, 1993). They differ from usual approaches to feed-forward neural networks in that they engage in a full Bayesian treatment, supplying a complete posterior distribution of forecasts (which are jointly Gaussian). For regression, they are also computationally relatively simple to implement, the basic model requiring only solving a system of linear equations, albeit one of size equal to the number of training examples (requiring $O(N^3)$ computation).

Moreover, a deep connection exists between Gaussian Processes and neural networks: it can be shown that the prior distribution over functions implied by a (Bayesian) one-layer feed-forward neural network tends to a Gaussian Process when the number of hidden units in the network tends to infinity, if a standard isotropic prior over network weights is assumed (Neal, 1996).

Motivation

The motivation for this work comes from forecasting and actively trading price spreads between commodity futures contracts (see, e.g., Hull 2005, for an introduction). Since futures contracts expire and have a finite duration, this problem is characterized by the presence of a large number of separate historical time series, which all can be of relevance in forecasting a new time series. For example, we expect seasonalities to affect similarly all the series. Furthermore, conditioning information, in the form of macroeconomic variables, can be of importance, but exhibit the cumbersome property of being released periodically, with explanatory power that varies across the forecasting horizon. In other words, when making a very long-horizon forecast, the model should not incorporate conditioning information in the same way as when making a short- or medium-term forecast. A possible solution to this problem is to have multiple models for forecasting each time series, one for each time scale. However, this is hard to work with, requires a high degree of skill on the part of the modeler, and is not amenable to robust automation when one wants to process hundreds of time series. In addition, in order to measure risk associated with a particular trade (buying at time t and selling at time t'), we need to estimate the *covariance of the price predictions* associated with these two points in the trajectory.

These considerations motivate the use of Gaussian processes, which naturally provide a covariance matrix between forecasts made at several points. To tackle the challenging task of forecasting and trading spreads between commodity futures, we introduce here a form of functional data analysis in which the function to be forecast is indexed both by the date of availability of the information set and by the forecast horizon. The predicted trajectory is thus represented as a functional object associated with a distribution,

a Gaussian process, from which the risk of different trading decisions can readily be estimated. This approach allows incorporating input variables that cannot be assumed to remain constant over the forecast horizon, like statistics of the short-term dynamics.

As an added benefit, the proposed method turns out to be very intuitive to practitioners. The notion of seeing a forecast trajectory for the spread fits much better with the way that traders work in managing trades, and seems more natural than next-period asset expected return and variance of traditional portfolio theory.[1]

Previous Work

Gaussian processes for time-series forecasting have been considered before. Multi-step forecasts are explicitly tackled by Girard et al. (2003), wherein uncertainty about the intermediate values is formally incorporated into the predictive distribution to obtain more realistic uncertainty bounds at longer horizons. However, this approach, while well-suited to purely autoregressive processes, does not appear amenable to the explicit handling of exogenous input variables. Furthermore, it suffers from the restriction of only dealing with regularly-sampled data. Our approach draws from the CO_2 model of Rasmussen and Williams (2006) as an example of application-specific covariance function engineering.

[1] Questions that are of daily concern to a trader go beyond the immediate sign of the position — whether *long* or *short* — but cover the validity of the entry and exit points, the expected profit from the trade, the expected timeframe and the conditions for which one should consider an early exit.

2. On Commodity Spreads

Until the work of Working (1949), it had been the norm to consider futures contracts at different maturities on the same underlying commodity as being "substantially independent", the factors impacting one contract (such as expectations of a large harvest in a given month) having little bearing on the others. His *theory of price of storage*, backed by a large body of empirical studies on the behavior of wheat futures (such as, e.g. Working 1934), invalidated the earlier views and established the basis of *interporal pricing relationships* that form a key element in understanding the spread behavior of storable commodities.

It has long been recognized that commodity prices, including spreads, can exhibit complex behavior. As Working aptly wrote more than 70 years ago (Working, 1935),

> If the important factors bearing on the price of any one commodity were always very few in number and related to the price in very simple fashion, direct multiple correlation analysis might appear entirely adequate, even in light of present general knowledge of its limitations. But intensive realistic study has revealed that, for some commodity prices, the number of factors that must be regarded as really important is rather large. Regressions are frequently curvilinear. The effects of price factors are often not independent, but joint. The factors, or at least the most suitable measures of them, are not known in advance, but remain to be determined; the character of the functional relationships between the factors, separately or jointly, and the price, is unknown and may not safely be assumed linear.

Kim and Leuthold (2000) provide evidence of large shifts in the distributional behavior of corn, live cattle, gold and T-bonds across time periods and temporal aggregation horizon.

Several authors have examined the specific influences of spread seasonalities and other eventual forecastable behavior. Simon (1999) finds evidence of a long-run equilibrium (cointegration) relationship in the soybeans crush spread,[2] once seasonality and linear trend are accounted for. He also shows mean-reversion at a five-day horizon. Simple (in-sample) trading rules that account for both phenomena show profit after transaction costs. Girma and Paulson (1998) find significant seasonality at both the monthly and trading-week levels in the petroleum complex spreads[3] and apply trading rules based on the observed seasonality to extract abnormal out-of-sample returns for the $3:2:1$ crack spread, less so for the simpler spreads. Dutt et al. (1997) examined the effects of intra- versus inter-crop year spreads on the volatility of agricultural futures spreads. In the precious metal spreads, Liu and Chou (2003) considered long-run parity relationships between the gold and silver markets and obtain that significant riskless profits could be earned based on the forecasts of an error-correcting model. This market was previously investigated by Wahab et al. (1994), and pure gold spreads were the subject of a study by Poitras (1987). In equity index futures, Butterworth and Holmes (2002) finds some intermarket mispricings between FTSE 100 and FTSE Mid 250 futures, although the possibilities of profitable trading after transaction costs appeared slim. More recently, Dunis et al. (2006b) applied

[2]Which consists in taking a long position in the soybeans contract, and offsetting shorts in both soybean meal and soybean oil

[3]Collectively referred to as "crack spreads".

nonlinear methods, such as recurrent neural networks and filter rules, to profitably trade petroleum spreads.

In this paper we shall consider the simplest type of calendar spreads: intracommodity calendar spreads, obtained by simultaneously taking a long and equal (in terms of the number of contracts) short position in two different maturities of the same underlying commodity. The price of the spread is simply given by the subtraction of the the two prices, since they express the same deliverable quantity. Moreover, many of the above papers introducing trading simulations consider the *continuous-contract* spread formed by rolling from one futures contract to another as they reach maturity. As shall be made clear below, a unique property of the methodology proposed in this paper is the ability to directly consider the separate histories of the spreads of interest in previous years and use them to forecast the evolution of a spread for the current year. As such, we can dispense with having to specify a rollover policy.

Organization of this paper

We start with a review of the main results pertaining to Gaussian processes for nonlinear regression (§3/p. 9), then explain in detail the methodology that we developed for forecasting the complete future trajectory of a spread using Gaussian processes (§4/p. 17). We follow with an explanation of the experimental setting for evaluating forecasting performance (§5/p. 26), and an account of the forecasting results of the proposed methodology against several benchmark models (§6/p. 35). We continue with the elaboration of a criterion to take price trajectory forecasts and turn them in trading decisions (§7/p. 35) and financial performance results on a portfolio of 30 spreads (§8/p. 40). Finally, §9/p. 46 presents directions for future

research.

3. Review of Gaussian Processes for Regression

The present section briefly reviews Gaussian processes at a level sufficient for understanding the spread forecasting methodology developed in the next section.

3.1. Basic Concepts for the Regression Case

A Gaussian process is a generalization of the Gaussian distribution: it represents a probability distribution over *functions* which is entirely specified by a mean and covariance *functions*. Borrowing the succinct definition from Rasmussen and Williams (2006), we formally define a Gaussian process (GP) as

Definition 1. *A* Gaussian process *is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Let $\mathbf{x}$ index into the real process $f(\mathbf{x})$.[4] We write

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\cdot), k(\cdot,\cdot)\right), \tag{1}$$

where functions $m(\cdot)$ and $k(\cdot,\cdot)$ are, respectively, the mean and covariance functions:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x})\right], \\ k(\mathbf{x}_1, \mathbf{x}_2) &= \mathbb{E}\left[\left(f(\mathbf{x}_1) - m(\mathbf{x}_1)\right)\left(f(\mathbf{x}_2) - m(\mathbf{x}_2)\right)\right]. \end{aligned}$$

[4] Contrarily to many treatments of stochastic processes, there is no necessity for $\mathbf{x}$ to represent time; indeed, in our application of Gaussian processes, some of the elements of $\mathbf{x}$ have a temporal interpretation, but most are general macroeconomic variables used to condition the targets.

In a regression setting, we shall assume that we are given a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}$, with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. For convenience, let $\mathbf{X}$ be the matrix of all training inputs, and $\mathbf{y}$ the vector of targets. We shall further assume that the y_i are noisy measurements from the underlying process $f(\mathbf{x})$:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_n^2).$$

Regression with a GP is achieved by means of Bayesian inference in order to obtain a posterior distribution over functions given a suitable prior and training data. Then, given new *test inputs*, we can use the posterior to arrive at a *predictive distribution* conditional on the test inputs and the training data. The predictive distribution is normal for a GP. Although the idea of manipulating distributions over functions may appear cumbersome, the *consistency property* of GPs means that any finite number of values sampled from the process f are jointly normal; hence inference over random functions can be shown to be completely equivalent to inference over a finite number random variables.

It is often convenient, for simplicity, to assume that the GP prior distribution has a mean of zero,

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\cdot, \cdot)).$$

Let $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N]$ be the vector of (latent) function values at the training inputs. Their prior distribution is given by

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K(\mathbf{X}, \mathbf{X})),$$

where $K(\mathbf{X}, \mathbf{X})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ is matrix formed by evaluating the covariance function between all pairs of training points. (This matrix is also known as

the kernel or Gram matrix.) We shall consider the joint prior distribution between the training and an additional set of *test points*, with locations given by the matrix $\mathbf{X}_*$, and whose function values $\mathbf{f}_*$ we wish to infer. Under the GP prior, we have[5]

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right). \tag{2}$$

We obtain the predictive distribution at the test points as follows. From Bayes' theorem, the joint posterior given the training data is

$$\begin{aligned} P(\mathbf{f}, \mathbf{f}_* \,|\, \mathbf{y}) &= \frac{P(\mathbf{y} \,|\, \mathbf{f}, \mathbf{f}_*) P(\mathbf{f}, \mathbf{f}_*)}{P(\mathbf{y})} \\ &= \frac{P(\mathbf{y} \,|\, \mathbf{f}) P(\mathbf{f}, \mathbf{f}_*)}{P(\mathbf{y})}, \end{aligned} \tag{3}$$

where $P(\mathbf{y} \,|\, \mathbf{f}, \mathbf{f}_*) = P(\mathbf{y} \,|\, \mathbf{f})$ since the likelihood is conditionally independent of $\mathbf{f}_*$ given $\mathbf{f}$, and

$$\mathbf{y} \,|\, \mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 I_N) \tag{4}$$

with I_N the $N \times N$ identity matrix. The desired predictive distribution is obtained by marginalizing out the training set latent variables,

$$\begin{aligned} P(\mathbf{f}_* \,|\, \mathbf{y}) &= \int P(\mathbf{f}, \mathbf{f}_* \,|\, \mathbf{y}) d\mathbf{f} \\ &= \frac{1}{P(\mathbf{y})} \int P(\mathbf{y} \,|\, \mathbf{f}) P(\mathbf{f}, \mathbf{f}_*) d\mathbf{f}. \end{aligned}$$

Since these distributions are all normal, the result of the (normalized) marginal is also normal, and can be shown to have mean and covariance given by

$$\mathbb{E}\left[\mathbf{f}_* \,|\, \mathbf{y}\right] = K(\mathbf{X}_*, \mathbf{X}) \boldsymbol{\Lambda}^{-1} \mathbf{y}, \tag{5}$$

$$\mathrm{Cov}\left[\mathbf{f}_* \,|\, \mathbf{y}\right] = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}) \boldsymbol{\Lambda}^{-1} K(\mathbf{X}, \mathbf{X}_*), \tag{6}$$

[5]Note that the following expressions in this section are implicitly conditioned on the training and test inputs, respectively $\mathbf{X}$ and $\mathbf{X}_*$. The explicit conditioning notation is omitted for brevity.

where we have set

$$\mathbf{\Lambda} = K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I_N. \tag{7}$$

Computation of $\mathbf{\Lambda}^{-1}$ is the most computationally expensive step in Gaussian process regression, requiring $O(N^3)$ time and $O(N^2)$ space.

Note that a natural outcome of GP regression is an expression for not only the expected value at the test points (5), but also the full covariance matrix between those points (6). We shall be making use of this covariance matrix later.

3.2. Choice of Covariance Function

We have so far been silent on the form that should take the covariance function $k(\cdot, \cdot)$. A proper choice for this function is important for encoding prior knowledge about the problem; several striking examples are given in Rasmussen and Williams (2006). In order to yield valid covariance matrices, the covariance function should, at the very least, be symmetric and positive semi-definite, which implies that all its eigenvalues are nonnegative,

$$\int k(\mathbf{u}, \mathbf{v}) f(\mathbf{u}) f(\mathbf{v}) \, d\mu(\mathbf{u}) \, d\mu(\mathbf{v}) \geq 0,$$

for all functions f defined on the appropriate space and measure μ.

Two common choices of covariance functions are the *squared exponential*[6],

$$k_{\mathrm{SE}}(\mathbf{u}, \mathbf{v}; \sigma_\ell) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|^2}{2\sigma_\ell^2}\right) \tag{8}$$

and the *rational quadratic*

$$k_{\mathrm{RQ}}(\mathbf{u}, \mathbf{v}; \sigma_\ell, \alpha) = \left(1 + \frac{\|\mathbf{u} - \mathbf{v}\|^2}{2\alpha\sigma_\ell^2}\right)^{-\alpha}. \tag{9}$$

[6] Also known as the Gaussian or radial basis function kernel.

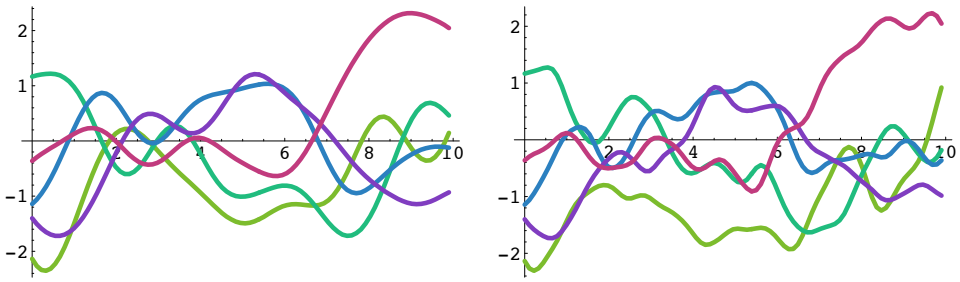

Figure 1: **Left:** Random functions drawn from the GP prior with the **squared exponential** covariance ($\sigma_\ell = 1$). **Right:** "Same functions" under the **rational quadratic** covariance ($\sigma_\ell = 1, \alpha = \frac{1}{2}$).

In both instances, the hyperparameter σ_ℓ governs the *characteristic length-scale* of the covariance function, indicating the degree of smoothness of the underlying random functions. The rational quadratic can be interpreted as an infinite mixture of squared exponentials with different length-scales; it converges to a squared exponential with characteristic length-scale σ_ℓ as $\alpha \to \infty$.

Figure 1 illustrates several functions drawn from the GP prior, respectively with the squared exponential and rational quadratic covariance functions. The same random numbers have served for generating both sets, so the functions are "alike" in some sense. Observe that a function under the squared exponential prior is smoother than the corresponding function under the rational quadratic prior.

3.3. Optimization of Hyperparameters

Most covariance functions, including those presented above, have free parameters — termed *hyperparameters* — that control their shape, for instance the characteristic length-scale σ_n^2. It remains to explain how their value should be set. This is an instance of the general problem of *model*

selection. For many machine learning algorithms, this problem has often been approached by minimizing a validation error through cross-validation (Stone, 1974) and such methods have been proposed for Gaussian processes (Sundararajan and Keerthi, 2001).

An alternative approach, quite efficient for Gaussian processes, consists in maximizing the *marginal likelihood*[7] of the observed data with respect to the hyperparameters. This function can be computed by introducing latent function values that are immediately marginalized over. Let θ the set of hyperparameters that are to be optimized; and let $K_{\mathbf{X}}(\theta)$ the covariance matrix computed by a covariance function whose hyperparameters are θ,

$$K_{\mathbf{X}}(\theta)_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j; \theta).$$

The marginal likelihood (here making explicit, for clarity, the dependence on the training inputs and hyperparameters) can be written

$$p(\mathbf{y} \mid \mathbf{X}, \theta) = \int p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) p(\mathbf{f} \mid \mathbf{X}, \theta)\, d\mathbf{f}, \tag{10}$$

where the distribution of observations $p(\mathbf{y} \mid \mathbf{f}, \mathbf{X})$, given by eq. (4), is conditionally independent of the hyperparameters given the latent function $\mathbf{f}$. Under the Gaussian process prior (see eq. (2)), we have $\mathbf{f} \mid \mathbf{X}, \theta \sim \mathcal{N}(0, K_{\mathbf{X}}(\theta))$, or in terms of log-likelihood,

$$\log p(\mathbf{f} \mid \mathbf{X}, \theta) = -\frac{1}{2}\mathbf{f}' K_{\mathbf{X}}^{-1}(\theta)\mathbf{f} - \frac{1}{2}\log |K_{\mathbf{X}}(\theta)| - \frac{N}{2}\log 2\pi. \tag{11}$$

Since both distributions in eq. (10) are normal, the marginalization can be carried out analytically to yield

$$\log p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2}\mathbf{y}'(K_{\mathbf{X}}(\theta) + \sigma_n^2 I_N)^{-1}\mathbf{y} - \frac{1}{2}\log \left| K_{\mathbf{X}}(\theta) + \sigma_n^2 I_N \right| - \frac{N}{2}\log 2\pi. \tag{12}$$

[7] Also called the *integrated likelihood* or (Bayesian) *evidence*.

This expression can be maximized numerically, for instance by a conjugate gradient algorithm (e.g. Bertsekas 2000) to yield the selected hyperparameters:

$$\theta^* = \arg\max_{\theta} \log p(\mathbf{y} \mid \mathbf{X}, \theta).$$

The likelihood function is in general non-convex and this maximization only finds a local maximum in parameter space; however, empirically it usually works very well for a large class of covariance functions, including those covered in §3.2/p. 12.

It should be mentioned that completely a Bayesian treatment would not be satisfied with such an optimization, since it only picks a single value for each hyperparameter. Instead, one would define a prior distribution on hyperparameters (an *hyperprior*), $p(\theta)$, and marginalize with respect to this distribution. However, for many "reasonable" hyperprior distributions, this is computationally expensive, and often yields no marked improvement over simple optimization of the marginal likelihood (MacKay, 1999).

The gradient of the marginal log-likelihood with respect to the hyperparameters — necessary for numerical optimization algorithms — can be expressed as

$$\frac{\partial \log p(\mathbf{y} \mid \mathbf{X}, \theta)}{\partial \theta_i} = \frac{1}{2}\mathbf{y}' K_{\mathbf{X}}^{-1}(\theta) \frac{\partial K_{\mathbf{X}}(\theta)}{\partial \theta_i} K_{\mathbf{X}}^{-1}(\theta)\mathbf{y} - \frac{1}{2}\mathrm{Tr}\left(K_{\mathbf{X}}^{-1}(\theta) \frac{\partial K_{\mathbf{X}}(\theta)}{\partial \theta_i} \right). \tag{13}$$

See Rasmussen and Williams (2006) for details on the derivation of this equation.

Two-Stage Training. Since hyperparameter optimization involves a large number of repetitions of solving a linear system, it requires a significantly greater computational effort than the single solution of eq. (5) and (6). To make this optimization tractable, yet not relinquish our ability to consider

enough historical data to obtain a model with adequate predictive power, the experiments of this paper (§5/p. 26) rely on a *two-stage training* procedure:

1. First, hyperparameters are optimized on a moderate-sized training set ($N = 500$).
2. Then, keeping hyperparameters fixed, we train the final model with a larger training-set ($N = 2000$–3000); our experiments, described below, used $N = 2250$.

3.4. Weighting the Importance of Input Variables: Automatic Relevance Determination

The covariance functions introduced in eq. (8) and (9) rely on an isotropic Euclidian norm as the similarity measure between two vectors in input space. These functions assume that a global characteristic length-scale σ_ℓ^2 governs proximity evaluation in all input dimensions. Even after normalizing the input variables to the same scale[8], differing predictive strength and noise level among the input variables may make it compelling to use a *specific characteristic length-scale* for each input.

This amounts to a simple rewriting of the Euclidian norm in the previously-introduced covariance functions as weighted norms, with a weight $\frac{1}{2}\sigma_i^2, i = 1, \ldots, D$ associated to each input dimension. For instance, the squared exponential kernel (eq. 8) would take the form

$$k_{\text{SE ARD}}(\mathbf{u}, \mathbf{v}; \sigma) = \exp\left(-\sum_{i=1}^{D} \frac{(\mathbf{u}_i - \mathbf{v}_i)^2}{2\sigma_i^2}\right). \tag{14}$$

[8] As can be done, for instance, by standardization, i.e. subtracting the mean of each variable and dividing by its standard deviation; see §4.3/p. 23.

The hyperparameters σ_i are found by numerically maximizing the marginal likelihood of the prior, as described in the previous section. After optimization, inputs that are found to have little importance are given a high σ_i, so that their importance in the norm is diminished. This procedure is an application to Gaussian processes of *automatic relevance determination*, a soft input variable selection procedure originally proposed in the context of Bayesian neural networks (MacKay, 1994; Neal, 1996). We revisit this topic in §4.5/p. 25 when describing the specific form of the covariance function used in our forecasting methodology.

4. Forecasting Methodology

We consider a set of N real time series each of length M_i, $\{y_t^i\}, i = 1, \ldots, N$ and $t = 1, \ldots, M_i$. In our application each i represents a different year, and the series is the sequence of commodity spread prices during the period where it is traded. The lengths of all series are not necessarily identical, but we shall assume that the time periods spanned by the series are "comparable" (e.g. the same range of days within a year if the series follow an annual cycle) so that knowledge from past series can be transferred to a new one to be forecast.

Note that there is no attempt to attempt to represent the whole history as a single continuous time series. Rather, each "trading year" of data[9] is treated as a separate time series,[10] and the year number is used as a con-

[9] Which may cross calendar years depending on the maturity months of the spread being modeled.

[10] Nothing in this treatment precludes individual spread price trajectories to span a longer than one year horizon, but for simplicity we shall use the "year" as the unit of trajectory length in the proceeds.

tinuous independent variable in the regression model. This representation is natural for spreads whose existence, like those of the underlying futures contracts, is tied to specific delivery months, and whose behavior intimately driven (for agricultural commodities) by seasonalities such as the prevailing crop conditions in a given year. A synthetic illustration of the data representation is shown in Figure 2. This should be contrasted to some spread modeling methodologies (e.g. Dunis et al. 2006a) that attempt to create a *continuous spread series* by defining a rollover policy over futures contracts.

The **forecasting problem** is that given observations from the complete series $i = 1, \ldots, N-1$ and from a *partial last series*, $\{y_t^N\}, t = 1, \ldots, M_N$, we want to extrapolate the last series until a predetermined endpoint, i.e. characterize the joint distribution of $\{y_\tau^N\}, \tau = M_N+1, \ldots, M_N+H$. We are also given a set of non-stochastic explanatory variables specific to each series, $\{\mathbf{x}_t^i\}$, where $\mathbf{x}_t^i \in \mathbb{R}^d$. Our objective is to find an effective representation of $P(\{y_\tau^N\}_{\tau=M_N+1,\ldots,M_N+H} \mid \{\mathbf{x}_t^i, y_t^i\}_{t=1,\ldots,M_i}^{i=1,\ldots,N})$, with τ, i and t ranging, respectively over the forecasting horizon, the available series and the observations within a series.

4.1. Functional Representation for Forecasting

In the spirit of functional data analysis, a first attempt at solving the spread evolution forecasting problem is to formulate it as one of regression from a "current date" (along with additional exogenous variables) to spread price. Contrarily to a most traditional stationary time-series model that would represent a finite-horizon series return, we include a *representation of the current date as an independent variable*, and regress on (appropriately normalized) spread prices. More specifically, we split the date input into two parts: the current time-series identity i (an integer representing, e.g., the

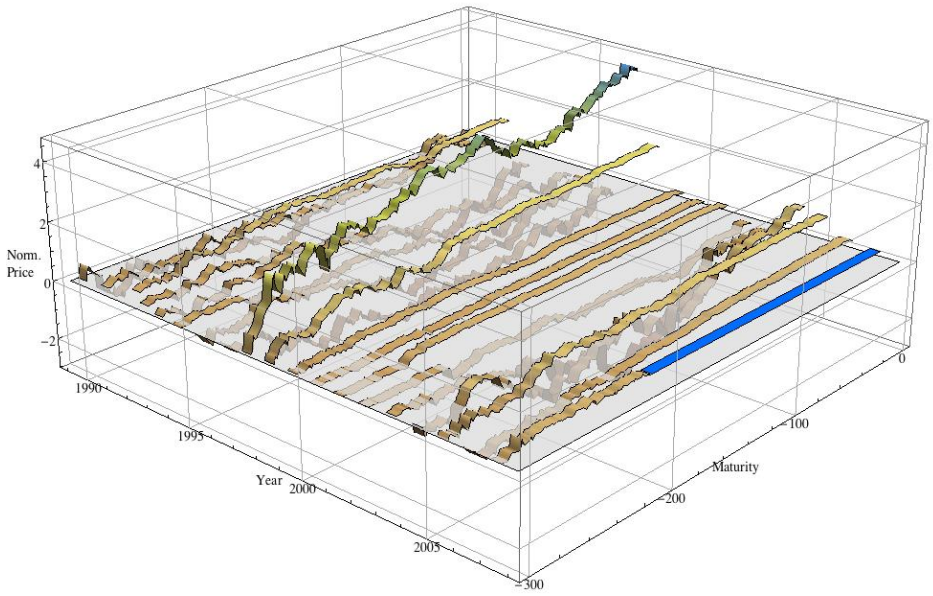

Figure 2: Illustration of the regression variables around which the forecasting problem is specified. Shown is the price history of the March–July old-crop/new-crop Wheat spread, as a function of the crop year and number of days to maturity. Prices are normalized to start at zero at maturity -300 and scaled to unit standard deviation over the sample. To aid the interpretation of price movement, prices that fall below the cutting plane at zero are shown slightly shaded ("under water"). The objective of the forecasting model is to fill out the "blue strip" in the last year of data, given the partial trajectory observed so far for that last year, and previous years complete trajectories.

spread trading year) and the time t within the series (we use and number of calendar days remaining until spread maturity, where maturity is defined as that of the near-term spread leg). This yields the model

$$\begin{aligned} \mathbb{E}\left[y_t^i \middle| \mathcal{I}_{t_0}^i\right] &= f(i, t, \mathbf{x}_{t|t_0}^i) \\ \mathrm{Cov}\left[y_t^i, y_{t'}^{i'} \middle| \mathcal{I}_{t_0}^i\right] &= g(i, t, \mathbf{x}_{t|t_0}^i, i', t', \mathbf{x}_{t'|t_0}^{i'}), \end{aligned} \tag{15}$$

these expressions being conditioned on the *information set* $\mathcal{I}_{t_0}^i$ containing information up to time t_0 of series i (we assume that all prior series $i' < i$ are also included in their entirety in $\mathcal{I}_{t_0}^i$). The notation $\mathbf{x}_{t|t_0}^i$ denotes a forecast of $\mathbf{x}_t^i$ given information available at t_0. Functions f and g result from Gaussian process training, eq. (5) and (6), using information in $\mathcal{I}_{t_0}^i$. To extrapolate over the unknown horizon, one simply evaluates f and g with the series identity index i set to N and the time index t within a series ranging over the elements of τ (forecasting period). Owing to the smoothness properties of an adequate covariance function, one can expect the last time series (whose starting portion is present in the training data) to be smoothly extended, with the Gaussian process borrowing from prior series, $i < N$, to guide the extrapolation as the time index reaches far enough beyond the available data in the last series. Figure 3 illustrates the approach. Note that we have additional input variables (on top of current year, maturity, and maturity delta described below), which are detailed in section 5.

The principal difficulty with this method resides in handling the exogenous inputs $\mathbf{x}_{t|t_0}^N$ over the forecasting period: the realizations of these variables, $\mathbf{x}_t^N$, are not usually known at the time the forecast is made and must be extrapolated with some reasonableness. For slow-moving variables that represent a "level" (as opposed to a "difference" or a "return"), one can conceivably keep their value constant to the last known realization across the

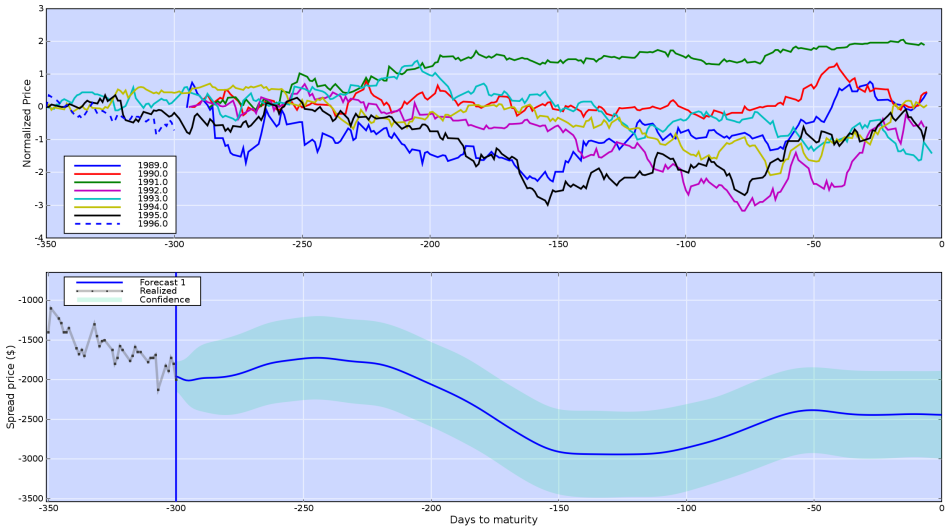

Figure 3: **Top part:** Training set for March–July Wheat spread, containing data until 1995/05/19; each spread year constitutes a distinct trajectory. The x-axis is the negative number of days to maturity (so that time flows to the right). The y-axis is normalized spread price, where the normalization procedure is described in the text. The partial trajectory observed so far for the current year (1996) is the dashed line. **Bottom part:** Forecasts made by extending the partial spread trajectory of the current year using the gaussian process regression methodology. The y-axis is the actual spread price in \$ (where individual contract prices are multiplied by a contract size of 50 for Wheat).

forecasting period. However, this solution is restrictive, problem-dependent, and precludes the incorporation of short-term dynamics variables (e.g. the first differences over the last few time-steps) if desired.

4.2. Augmented Functional Representation

We propose in this paper to augment the functional representation with an additional input variable that expresses the time *at which* the forecast is being made, in addition to the time *for which* the forecast is made. We shall denote the former the *operation time* and the latter the *target time*.

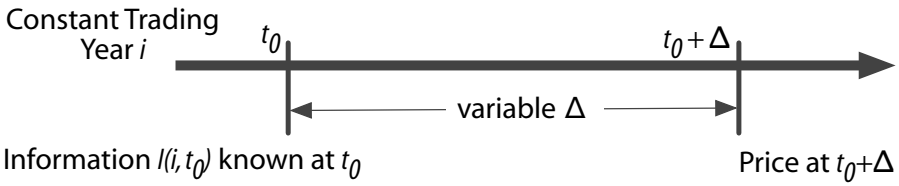

Figure 4: The augmented functional representation attempts to capture the relationship between the price at the target time $t_0 + \Delta$ and information available at the operation time t_0, for all t_0 and Δ within a given spread trading year.

The distinction is as follows: **operation time** represents the time at which the other input variables are observed and the time at which, conceptually, a forecast of the entire future trajectory is performed. In contrast, **target time** represents time at a point of the predicted target series (beyond operation time), given the information known at the operation time. Figure 4 illustrates the concept.

The training set used to form the augmented representation conceptually includes all pairs $\langle \textit{operation time}, \textit{target time} \rangle$ for a given trajectory. As previously, the time series index i remains part of the inputs. In this framework, forecasting is performed by holding the time series index constant to N, the operation time constant to the time M_N of the last observation, the other input variables constant to their last-observed values $\mathbf{x}^N_{M_N}$, and *varying the target time* over the forecasting period τ. Since we are not attempting to extrapolate the inputs beyond their intended range of validity, this approach admits general input variables, without restriction as to their type, and whether they themselves can be forecast.

It can be convenient to represent the target time as a positive delta Δ from the operation time t_0. In contrast to eq. (16), this yields the represen-

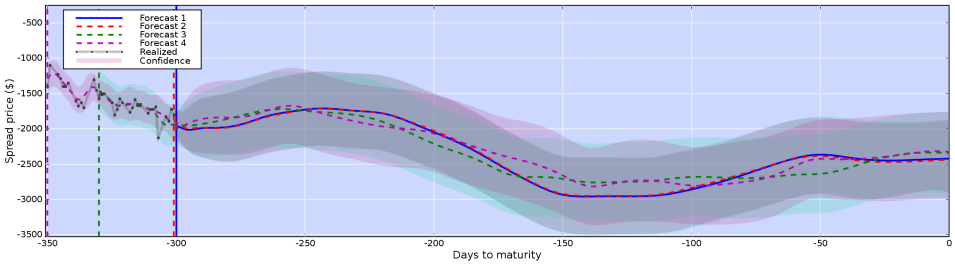

Figure 5: Forecasts made from several operation times, given the same training set as in Figure 3.

tation

$$\begin{aligned} \mathbb{E}\left[y^i_{t_0+\Delta}\middle|\mathcal{I}^i_{t_0}\right] &= f(i,t_0,\Delta,\mathbf{x}^i_{t_0}) \\ \mathrm{Cov}\left[y^i_{t_0+\Delta}, y^{i'}_{t'_0+\Delta'}\middle|\mathcal{I}^i_{t_0}\right] &= g(i,t_0,\Delta,\mathbf{x}^i_{t_0},i',t'_0,\Delta',\mathbf{x}^{i'}_{t'_0}), \end{aligned} \tag{16}$$

where we have assumed the operation time to coincide with the end of the information set. Note that this augmentation allows to dispense with the problematic extrapolation $\mathbf{x}^i_{t|t_0}$ of the inputs, instead allowing a direct use of the last available values $\mathbf{x}^i_{t_0}$. Moreover, from a given information set, nothing precludes forecasting the same trajectory from several operation times $t' < t_0$, which can be used as a means of evaluating the stability of the obtained forecast. This idea is illustrated in Figure 5.

4.3. Input and Target Variable Preprocessing

Input variables are subject to minimal preprocessing before being provided as input to the gaussian process: we standardize them to zero mean and unit standard deviation. The price targets require additional treatment: since the price level of a spread can vary significantly from year to year, we normalize the price trajectories to *start at zero* at the start of every year, by subtracting the first price. Furthermore, in order to get slightly better

behaved optimization, we divide the price targets by their overall standard deviation.

4.4. Training Set Subsampling

A major practical concern with gaussian processes is their ability to scale to larger problems. As seen in §3.1/p. 9, the basic training step of gaussian process regression involves the inversion of an $N \times N$ matrix (where N is the number of training examples), or the equivalent solution of a linear system. As such, training times can be expected to scale as $O(N^3)$. In practice, more than one such step is required: the optimization of hyperparameters described in §3.3/p. 13 involves the numerical optimization of a marginal likelihood function, requiring repeated solutions of same-sized problems — from several dozen to a few hundred times, depending on the number of hyperparameters and the accuracy sought.

The problem is compounded by augmentation, which requires, in principle, a greatly expanded training set. In particular, the training set must contain sufficient information to represent the output variable for the *many combinations of operation and target times* that can be provided as input. In the worst case, this implies that the number of training examples grows quadratically with the length of the training time series. In practice, a downsampling scheme is used wherein only a fixed number of target-time points is sampled for every operation-time point.[11]

A large number of approximation methods have been proposed to tackle large training sets; see Quiñonero-Candela and Rasmussen (2005) and Ras-

[11]This number was 15 in our experiments, and these were not regularly spaced, with longer horizons spaced farther apart. Furthermore, the original daily frequency of the data was reduced to keep approximately one operation-time point per week.

mussen and Williams (2006) for good surveys of the field. Most approaches involve some form of subsampling of the training set, used in conjunction with revised form of the estimators that can account for the influence of excluded training examples. Our downsampling scheme amounts to the simplest approximation method, called *subset of data*, which consists simply in taking a subset of the original examples.

4.5. Covariance Function Engineering

One of the most significant attractions of Gaussian processes lies in their ability to tailor their behavior for a specific application through the choice of covariance function. It is even possible to create completely novel functions, as long as they satisfy the positive semi-definiteness requirements of a valid covariance function (see §3.2/p. 12), or construct new ones according to straightforward composition rules (Shawe-Taylor and Cristianini, 2004; Rasmussen and Williams, 2006) — a process often called *covariance function (*or *kernel) engineering*.

In the present application, we used a modified form of the *rational quadratic* covariance function with hyperparameters for automatic relevance determination , which is expressed as

$$k_{\text{AUG-RQ}}(\mathbf{u}, \mathbf{v}; \ell, \alpha, \sigma_f, \sigma_{\text{TS}}) = \sigma_f^2 \left(1 + \frac{1}{2\alpha} \sum_{k=1}^{d} \frac{(\mathbf{u}_k - \mathbf{v}_k)^2}{\ell_k^2}\right)^{-\alpha} + \sigma_{\text{TS}}^2 \delta_{i_{\mathbf{u}}, i_{\mathbf{v}}}, \tag{17}$$

where $\delta_{j,k} \triangleq I[j = k]$ is the Kronecker delta. The variables $\mathbf{u}$ and $\mathbf{v}$ are values in the augmented representation introduced previously, containing the three variables representing time (current time-series index or year, operation time, target time) as well as the additional explanatory variables. The notation $i_{\mathbf{u}}$ denotes the time-series index component i of input variable $\mathbf{u}$.

The last term of the covariance function, the Kronecker delta, is used to induce an increased similarity among points that belong to the same time series (e.g. the same spread trading year). By allowing a series-specific average level to be maintained into the extrapolated portion, the presence of this term was found to bring better forecasting performance.

The hyperparameters $\ell_i, \alpha, \sigma_f, \sigma_{\text{TS}}, \sigma_n$ are found by maximizing the marginal likelihood on the training set by a standard conjugate gradient optimization, as outlined in §3.3/p. 13.

5. Experimental Setting

To establish the benefits of the proposed functional representation for forecasting commodity spread prices, we compared it against other likely models on three common grain and grain-related spreads:[12] the Soybeans (January–July, July–November, July–September, August–November, August–March), Soybean Meal (May–September July–December, July–September, August–December, August–March), and Soft Red Wheat (March–July, March–September, May–December, May–September, July–December). The forecasting task is to *predict the complete future trajectory* of each spread (taken individually), from 200 days before maturity until maturity.

[12] Our convention is to first give the *short leg* of the spread, followed by the *long leg*. Hence, Soybeans 1–7 should be interpreted as taking a short position (i.e. selling) in the January Soybeans contract and taking an offsetting long (i.e. buying) in the July contract. Traditionally, intra-commodity spread positions are taken so as to match the number of contracts on both legs — the number of short contracts equals the number of long ones — not the dollar value of the long and short sides.

5.1. Methodology

Realized prices in the previous trading years are provided from 250 days to maturity, using data going back to 1989. The first test year is 1994. Within a given trading year, the time variables represent the number of calendar days to maturity of the near leg; since no data is observed on week-ends, training examples are sampled on an irregular time scale.

Performance evaluation proceeds through a *sequential validation* procedure (Chapados and Bengio, 2001):[13] within a trading year, we first train models 200 days before maturity and obtain a first forecast for the future price trajectory. We then retrain models every 25 days, and obtain revised portions of the remainder of the trajectory. Proceeding sequentially, this operation is repeated for succeeding trading years. All forecasts are compared amongst models on squared-error and negative log-likelihood criteria (see "assessing significance", below).

5.2. Models Compared

The "complete" model to be compared against others is based on the augmented-input representation Gaussian process with the modified rational quadratic covariance function eq. (17). In addition to the three variables required for the representation of time, the following inputs were provided to the model: (i) the current spread price and the price of the three nearest futures contracts on the underlying commodity term structure, (ii) economic variables (the stock-to-use ratio and year-over-year difference in total ending stocks) provided on the underlying commodity by the U.S. Department of

[13] Also known as a "simulated out-of-sample forecasting methodology"; see, e.g. Stock and Watson (1999).

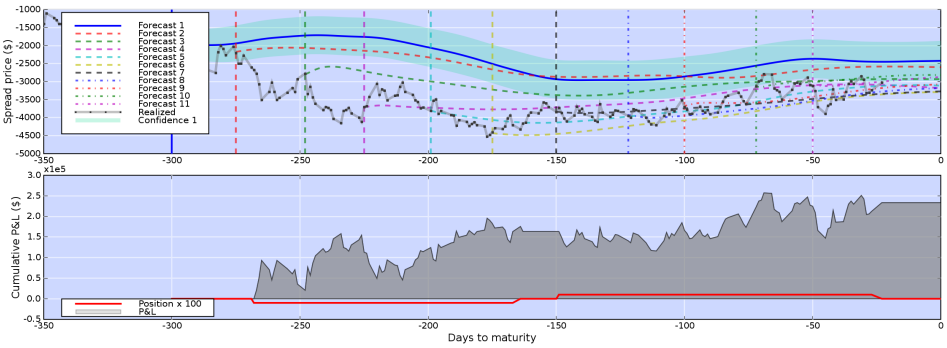

Figure 6: **Top Panel:** Illustration of multiple forecasts, repeated every 25 days, of the 1996 March–July Wheat spread (dashed lines); realized price is in gray. Although the first forecast (smooth solid blue, with confidence bands) mistakes the overall price level, it approximately correctly identifies local price maxima and minima, which is sufficient for trading purposes. **Bottom Panel:** Position taken by the trading model (in red: short, then neutral, then long), and cumulative profit of that trade (gray).

Agriculture (U.S. Department of Agriculture, 2008). This model is denoted **AugRQ/all-inp**. An example of the sequence of forecasts made by this model, repeated every 25 times steps, is shown in the upper panel of Figure 6.

To determine the value added by each type of input variable, we include in the comparison two models based on exactly on the same architecture, but providing less inputs: **AugRQ/less-inp** does not include the economic variables. **AugRQ/no-inp** further removes the price inputs, leaving only the time-representation inputs. Moreover, to quantify the performance gain of the augmented representation of time, the model **StdRQ/no-inp** implements a "standard time representation" that would likely be used in a functional data analysis model; as described in eq. (16), this uses a single time variable instead of splitting the representation of time between the operation and target times.

Finally, we compare against simpler models: **Linear/all-inp** uses a dot-

product covariance function to implement Bayesian linear regression, using the full set of input variables described above. And **AR**(**1**) is a simple linear autoregressive model. For this last model, the predictive mean and covariance matrix are established as follows (see, e.g. Hamilton 1994). We consider the scalar data generating process

$$y_t = \phi\, y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \tag{18}$$

where the process $\{y_t\}$ has an unconditional mean of zero.[14] The h-step ahead forecast from time t under this model, which we write $y_{t+h|t}$, is obtained by iterating h times eq. (18),

$$y_{t+h|t} = \phi^h y_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}. \tag{19}$$

The minimum mean-squared point forecast, $\hat{y}_{t+h|t}$, is given by the conditional expectation given information available at time t, $\mathcal{I}_t$,

$$\hat{y}_{t+h|t} = \mathbb{E}\left[y_{t+h} \,|\, \mathcal{I}_t\right] = \phi^h y_t. \tag{20}$$

The conditional variance of $y_{t+h|t}$ is given by

$$\text{Var}\left[y_{t+h} \,|\, \mathcal{I}_t\right] = \sum_{i=0}^{h-1} \phi^{2i}\sigma^2 = \sigma^2 \frac{\phi^{2h} - 1}{\phi^2 - 1}. \tag{21}$$

To evaluate the covariance between two forecasts, respectively at h and h' steps ahead, we start by evaluating the conditional expectation of the

[14] In practice, we subtract the empirical mean on the training set before proceeding with the analysis.

product,

$$\begin{aligned}
\mathbb{E}\left[y_{t+h|t}\, y_{t+h'|t} \,|\, \mathcal{I}_t\right] &= \mathbb{E}\left[\left(\phi^h y_t + \sum_{i=0}^{h-1} \phi^i \varepsilon_{t+h-i}\right)\left(\phi^{h'} y_t + \sum_{i=0}^{h'-1} \phi^i \varepsilon_{t+h'-i}\right)\middle|\, \mathcal{I}_t\right] \\
&= \phi^{h+h'} y_t^2 + \mathbb{E}\left[\sum_{i=0}^{h-1} \phi^{h-i-1} \varepsilon_{t+i} \sum_{j=0}^{h'-1} \phi^{h'-j-1} \varepsilon_{t+j}\middle|\, \mathcal{I}_t\right] \\
&= \phi^{h+h'} y_t^2 + \sigma^2 \sum_{i=0}^{M} \phi^{h+h'-2i-2} \\
&= \phi^{h+h'} y_t^2 + \sigma^2 \phi^{h+h'} \frac{1-\phi^{-2M}}{\phi^2-1},
\end{aligned}$$

with $M \triangleq \min(h, h')$. The covariance is given as

$$\begin{aligned}
\mathrm{Cov}\left[y_{t+h|t}, y_{t+h'|t} \,|\, \mathcal{I}_t\right] &= \mathbb{E}\left[y_{t+h|t}\, y_{t+h'|t} \,|\, \mathcal{I}_t\right] - \mathbb{E}\left[y_{t+h|t} \,|\, \mathcal{I}_t\right] \mathbb{E}\left[y_{t+h'|t} \,|\, \mathcal{I}_t\right] \\
&= \phi^{h+h'} y_t^2 + \sigma^2 \phi^{h+h'} \frac{1-\phi^{-2M}}{\phi^2-1} - \left(\phi^h y_t\right)\left(\phi^{h'} y_t\right) \\
&= \sigma^2 \phi^{h+h'} \frac{1-\phi^{-2M}}{\phi^2-1}. \qquad (22)
\end{aligned}$$

Equations (20) and (22) are used in conjunction with the approach outlined in §7/p. 35 to make trading decisions.

It should be noted that the iterated one-step linear forecaster is optimal at longer horizons only when the model is properly specified, i.e. the model matches the DGP[15] (Clements and Hendry, 1998). Otherwise, a direct forecasting at the desired horizon is preferable. However, due to complexity of estimating a large number of concurrent models in order to forecast the entire spread trajectory at all possible future horizons, and especially the computation of a reliable covariance matrix among them, we settled on using the iterated one-step model as our "naïve" benchmark.

[15] Data Generating Process.

5.3. Significance of Forecasting Performance Differences

For each trajectory forecast, we measure the squared error (SE) made at each time-step along with the negative log-likelihood (NLL) of the realized price under the predictive distribution. To account for differences in target variable distribution throughout the years, we normalize the SE by dividing it by the standard deviation of the test targets in a given year. Similarly, we normalize the NLL by subtracting the likelihood of a univariate Gaussian distribution estimated on the test targets of the year.

Due to the serial correlation it exhibits, the time series of performance differences (either SE or NLL) between two models cannot directly be subjected to a standard t-test of the null hypothesis of no difference in forecasting performance. The well-known Diebold-Mariano test (Diebold and Mariano, 1995) corrects for this correlation structure in the case where a *single time series* of performance differences is available. This test is usually expressed as follows. Let $\{d_t\}$ be the sequence of *error differences* between two models to be compared. Let $\bar{d} = \frac{1}{M}\sum_t d_t$ be the mean difference. The sample variance of $\bar{d}$ is readily shown (Diebold and Mariano, 1995) to be

$$\hat{v}_{\mathrm{DM}} \stackrel{\triangle}{=} \mathrm{Var}[\bar{d}] = \frac{1}{M}\sum_{k=-K}^{K}\hat{\gamma}_k,$$

where M is the sequence length and $\hat{\gamma}_k$ is an estimator of the lag-k autocovariance of the d_ts. The maximum lag order K is a parameter of the test and must be determined empirically. Then the statistic $DM = \bar{d}/\sqrt{\hat{v}_{\mathrm{DM}}}$ is asymptotically distributed as $\mathcal{N}(0,1)$ and a classical test of the null hypothesis $\bar{d} = 0$ can be performed.

Unfortunately, even the Diebold-Mariano correction for autocorrelation is not sufficient to compare models in the present case. Due to the repeated

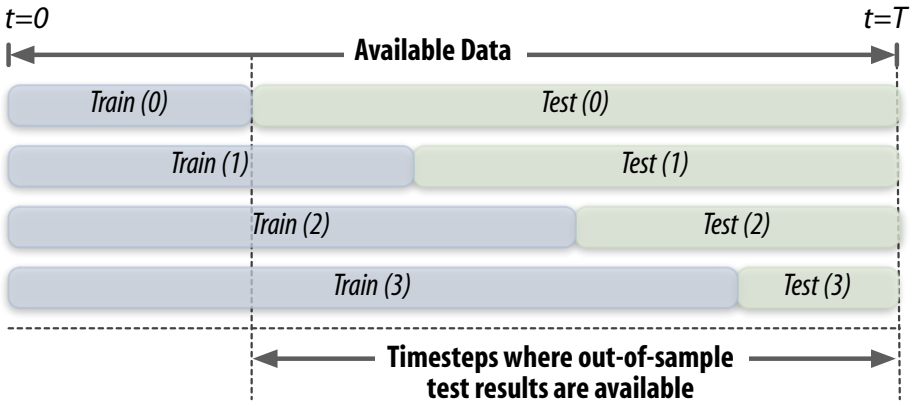

Figure 7: Illustration of **sequential validation** with stongly overlapping test sets. A model is retrained at regular intervals, each time tested on the remainder of the data. At each iteration, parts of the test data from the previous iteration is added to the training set.

forecasts made for the same time-step across *several iterations of sequential validation*, the error sequences are likely to be *cross-correlated* since they result from models estimated on strongly overlapping training sets. This effect is illustrated in Figure 7.

Although generally omitted, the steps in the derivation of the Diebold-Mariano variance estimator are useful to understand in order to generalize to the case of sequential validation. In order to perform an hypothesis test, we need an estimator of the variance of the sample mean in the case where the elements exhibit correlation. We have

$$\begin{aligned} \mathrm{Var}[\hat{d}] &= \frac{1}{N^2}\,\mathrm{Var}\left[\sum_i d_i\right] \\ &= \frac{1}{N^2}\sum_i\sum_j \mathrm{Cov}[d_i, d_j]. \end{aligned} \tag{23}$$

Diebold and Mariano work under the hypothesis of stationarity and maxi-

mum lag order of the covariances, such that

$$\mathrm{Cov}[d_i, d_j] = \begin{cases} \gamma_{|i-j|} & \text{if } |i-j| \leq K, \\ 0 & \text{if } |i-j| > K. \end{cases} \tag{24}$$

They then consider a *typical row* i of the covariance matrix, and notice that for this row, the inner summation in eq. (23) is equal to the sum of the covariance spectrum, $\sum_{k=-K}^{K} \gamma_k$. They then posit that this inner summation is repeated *for every row of the covariance matrix* (even on the boundaries), presumably to offset the fact that the higher-order correlations (greater than K) have been dropped. Substituting in (23), we have

$$\begin{aligned} \mathrm{Var}[\hat{d}] &= \frac{1}{N^2} \sum_i \sum_{k=-K}^{K} \gamma_k \\ &= \frac{1}{N} \sum_{k=-K}^{K} \gamma_k, \end{aligned}$$

where the last line follows from the inner summation not depending on the outer. This is the Diebold-Mariano variance estimator.

Generalization for Sequential Validation. For simplicity, we shall assume only two iterations of sequential validation, with the following elements resulting from testing. Test sets 1 and 2 overlap at timesteps 4 to 6:

Test set #1	1	2	3	4	5	6
Test set #2				4	5	6

Assuming covariance stationarity and keeping a maximum lag order $K = 2$, we obtain the following form for the covariance matrix linking all elements

$$\left(\begin{array}{ccc|ccc|ccc}
\gamma_0^1 & \gamma_1^1 & \gamma_2^1 & & & & & & \\
\gamma_1^1 & \gamma_0^1 & \gamma_1^1 & \gamma_2^1 & & & & & \\
\gamma_2^1 & \gamma_1^1 & \gamma_0^1 & \gamma_1^1 & \gamma_2^1 & & & & \\
 & \gamma_2^1 & \gamma_1^1 & \gamma_0^1 & \gamma_1^1 & \gamma_2^1 & \gamma_0^{1,2} & \gamma_1^{1,2} & \gamma_2^{1,2} \\
 & & \gamma_2^1 & \gamma_1^1 & \gamma_0^1 & \gamma_1^1 & \gamma_1^{1,2} & \gamma_0^{1,2} & \gamma_1^{1,2} \\
 & & & \gamma_2^1 & \gamma_1^1 & \gamma_0^1 & \gamma_2^{1,2} & \gamma_1^{1,2} & \gamma_0^{1,2} \\
\hline
 & & & \gamma_0^{1,2} & \gamma_1^{1,2} & \gamma_2^{1,2} & \gamma_0^2 & \gamma_1^2 & \gamma_2^2 \\
 & & & \gamma_1^{1,2} & \gamma_0^{1,2} & \gamma_1^{1,2} & \gamma_1^2 & \gamma_0^2 & \gamma_1^2 \\
 & & & \gamma_2^{1,2} & \gamma_1^{1,2} & \gamma_0^{1,2} & \gamma_2^2 & \gamma_1^2 & \gamma_0^2
\end{array}\right)$$

where γ_k^i denote the lag-k autocovariances within test set i, and $\gamma_k^{i,j}$ denote the lag-k cross-covariances between test sets i and j. The horizontal and vertical lines have been set between the elements belonging to test boundaries.

This matrix make explicit the impact of the cross-covariances (the off-diagonal blocks) the resulting variance. Our extension to the Diebold-Mariano test consists in simply estimating those terms and incorporating them into the overall variance estimator, yielding the *cross-covariance-corrected Diebold-Mariano variance estimator*

$$\hat{v}_{\text{CCC}-\text{DM}} = \frac{1}{M^2}\left(\sum_i M_i \sum_{k=-K}^{K} \hat{\gamma}_k^i + \sum_i \sum_{j \neq i} M_{i \cap j} \sum_{k=-K'}^{K'} \hat{\gamma}_k^{i,j}\right), \qquad (25)$$

where M_i is the number of examples in test set i, $M = \sum_i M_i$ is the total number of examples, $M_{i \cap j}$ is the number of time-steps where test sets i and j overlap, $\hat{\gamma}_k^i$ denote the estimated lag-k autocovariances within test set i, and $\hat{\gamma}_k^{i,j}$ denote the estimated lag-k cross-covariances between test sets i and j.

The maximum lag order for cross-covariances, K', is possibly different from K (our experiments used $K = K' = 15$). This revised variance estimator was used in place of the usual Diebold-Mariano statistic in the results presented below.

6. Forecasting Performance Results

Results of the forecasting *performance difference* between **AugRQ/all-inp** and all other models is shown in Table 1. We observe that **AugRQ/all-inp** generally beats the others on both the SE and NLL criteria, often statistically significantly so. In particular, the augmented representation of time is shown to be of value (i.e. comparing against **StdRQ/no-inp**). Moreover, the Gaussian process is capable of making good use of the additional price and economic input variables, although not always with the traditionally accepted levels of significance.

Figure 8 illustrates the same results graphically, and makes it easier to see, at a glance, how **AugRQ/all-inp** performs against other models.

A more detailed investigation of absolute and relative forecast errors of each model, on the March–July Wheat spread and for every year of the 1994–2007 period, is presented in appendix (§10/p. 47).

7. From Forecasts to Trading Decisions

We applied this forecasting methodology based on an augmented representation of time to trading a portfolio of spreads. Within a given trading year, we apply an information-ratio criterion to greedily determine the best trade into which to enter, based on the entire price forecast (until the end of the year) produced by the Gaussian process. More specifically, let $\{p_t\}$

Table 1: Forecast performance difference between **AugRQ/all-inp** and all other models, for the three spreads studied. For both the Squared Error and NLL criteria, the value of the cross-correlation-corrected statistic is listed (CCC) along with its p-value under the null hypothesis. A negative CCC statistic indicates that **AugRQ/all-inp** beats the other model on average.

	SoybeanMeal 5–9				**SoybeanMeal 7–12**				**SoybeanMeal 7–9**				**SoybeanMeal 8–12**				**SoybeanMeal 8–3**			
	Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL	
	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p
AugRQ/less-inp	−1.05	0.294	−0.94	0.344	−1.06	0.288	−1.23	0.217	−0.66	0.509	−1.26	0.207	−0.28	0.772	−1.08	0.277	0.46	0.644	0.29	0.765
AugRQ/no-inp	−1.78	0.075	−2.42	0.015	−2.13	0.032	−3.31	10^{-3}	−2.56	0.010	−3.09	0.001	−1.08	0.276	−1.66	0.096	−0.68	0.490	−2.30	0.021
Linear/all-inp	−1.61	0.107	−2.00	0.045	−2.16	0.030	−2.60	0.009	−1.73	0.083	−2.38	0.017	−1.84	0.065	−2.07	0.037	−1.48	0.138	−1.65	0.097
$AR(1)$	−2.53	0.011	0.12	0.904	−1.36	0.172	0.07	0.943	−1.78	0.074	−0.62	0.529	−1.62	0.104	−0.16	0.866	−3.82	10^{-4}	−2.68	0.007
StdRQ/all-inp	−0.68	0.496	0.323	0.747	−0.38	0.703	0.26	0.793	−1.06	0.290	−0.11	0.913	0.02	0.983	0.01	0.988	0.18	0.858	−2.48	0.013
StdRQ/no-inp	−2.69	0.007	−1.08	0.278	−1.37	0.169	0.22	0.822	−1.00	0.313	0.06	0.945	−1.56	0.118	−0.12	0.897	−3.03	0.002	−1.52	0.126
	Soybeans 1–7				**Soybeans 7–11**				**Soybeans 7–9**				**Soybeans 8–11**				**Soybeans 8–3**			
	Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL	
	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p
AugRQ/less-inp	−0.86	0.392	−0.89	0.372	−0.86	0.389	−1.32	0.183	−0.90	0.367	−1.75	0.079	−1.45	0.145	−1.70	0.087	−0.63	0.522	−1.29	0.195
AugRQ/no-inp	−1.68	0.093	−1.72	0.083	−2.93	0.003	−3.00	0.002	−1.91	0.056	−2.31	0.020	−1.55	0.119	−2.21	0.026	−0.29	0.769	−2.38	0.017
Linear/all-inp	−1.53	0.126	−1.32	0.183	−0.59	0.551	−1.36	0.172	−0.35	0.723	−1.24	0.213	−1.15	0.247	−1.84	0.065	−0.62	0.530	−1.69	0.090
$AR(1)$	−4.25	10^{-5}	−0.44	0.657	−1.54	0.121	0.17	0.859	−1.66	0.096	−0.44	0.654	−2.36	0.017	−1.18	0.236	−2.28	0.022	−0.13	0.894
StdRQ/all-inp	0.31	0.754	0.71	0.479	0.03	0.975	−0.47	0.641	0.40	0.689	−0.65	0.518	0.74	0.462	−1.33	0.184	0.93	0.352	−0.16	0.871
StdRQ/no-inp	−2.44	0.014	−1.02	0.304	−1.21	0.222	0.04	0.965	−0.72	0.467	−0.24	0.802	−0.99	0.319	−1.12	0.261	−1.65	0.097	−1.16	0.244
	Wheat 3–7				**Wheat 3–9**				**Wheat 5–12**				**Wheat 5–9**				**Wheat 7–12**			
	Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL		Sq. Error		NLL	
	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p	CCC	p
AugRQ/less-inp	−0.05	0.963	1.05	0.291	−0.28	0.778	−1.15	0.246	−0.67	0.497	−0.45	0.651	−0.41	0.678	−0.83	0.402	−0.24	0.808	−0.51	0.607
AugRQ/no-inp	−2.75	0.006	−2.42	0.015	−3.14	0.001	−3.26	0.001	−1.72	0.084	−2.26	0.023	−1.88	0.059	−2.54	0.010	−2.02	0.043	−1.53	0.125
Linear/all-inp	−4.20	10^{-5}	−3.45	10^{-4}	−2.36	0.017	−2.04	0.040	−2.07	0.038	−2.32	0.019	−1.88	0.058	−2.56	0.010	−0.76	0.444	−0.97	0.331
$AR(1)$	−6.50	10^{-9}	−6.07	10^{-9}	−5.95	10^{-9}	−4.38	10^{-5}	−4.88	10^{-6}	−5.55	10^{-8}	−4.55	10^{-6}	−4.55	10^{-6}	−3.76	10^{-4}	−2.30	0.021
StdRQ/all-inp	−0.65	0.514	−2.51	0.012	−0.49	0.628	−2.51	0.012	−1.56	0.125	−3.86	10^{-4}	−1.77	0.077	−3.23	0.001	−1.07	0.283	−0.61	0.543
StdRQ/no-inp	−2.67	0.007	−9.35	0.000	−2.41	0.015	−5.50	10^{-8}	−2.24	0.024	−4.36	10^{-5}	−2.12	0.033	−2.41	0.015	−1.42	0.153	−1.56	0.117

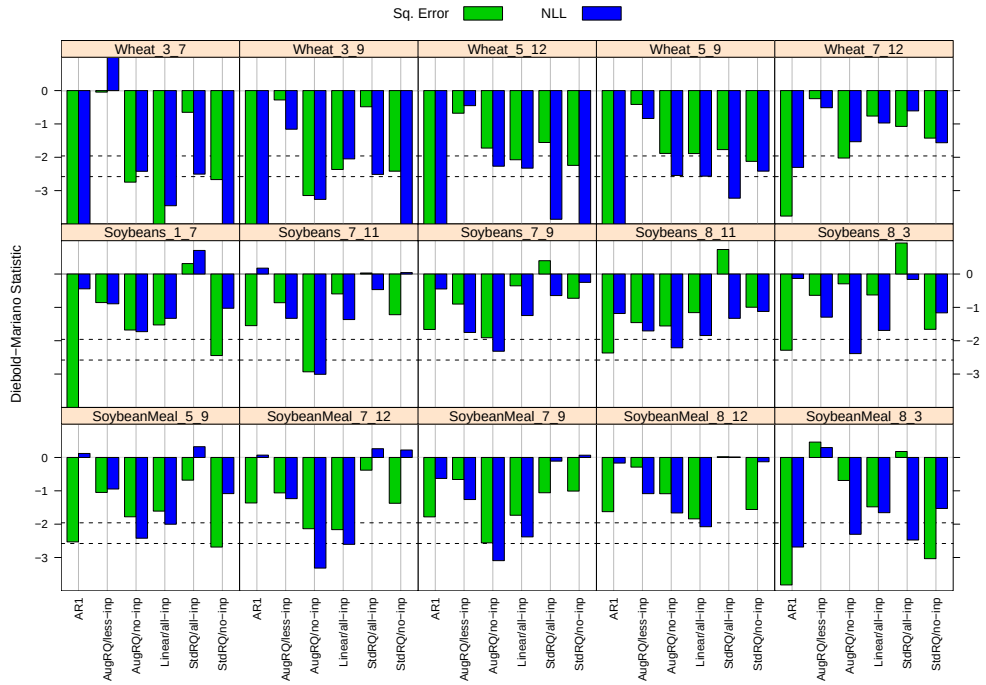

Figure 8: Overview of the Cross-Correlation-Corrected Diebold-Mariano test results for the grain spreads. The 1% and 5% thresholds for statistical significance (two-tailed test) are indicated by the dashed horizontal lines.

be the future prices forecast by the model at some operation time (presumably the time of last available element in the training set). The expected forecast dollar profit of buying at t_1 and selling at t_2 is simply given by $\tilde{p}_{t_2} - \tilde{p}_{t_1}$, where $\tilde{p}_{t_i}$ is the present-value discounted price: $\tilde{p}_{t_i} = e^{-r_f m_i} p_{t_i}$, with $m_i = t_i - t_0$ the number of days for which to discount and r_f a risk-free rate.

Of course, a prudent investor would take trade risk into consideration. A simple approximation of risk is given by the trade profit volatility. This yields the *forecast information ratio*[16] of the trade

$$\widehat{IR}(t_1, t_2) = \frac{\mathbb{E}[\tilde{p}_{t_2} - \tilde{p}_{t_1} | \mathcal{I}_{t_0}]}{\sqrt{\mathrm{Var}[\tilde{p}_{t_2} - \tilde{p}_{t_1} | \mathcal{I}_{t_0}]}}, \tag{26}$$

where $\mathrm{Var}[\tilde{p}_{t_2} - \tilde{p}_{t_1} | \mathcal{I}_{t_0}]$ can be computed as $\mathrm{Var}[\tilde{p}_{t_1} | \mathcal{I}_{t_0}] + \mathrm{Var}[\tilde{p}_{t_1} | \mathcal{I}_{t_0}] - 2\,\mathrm{Cov}[\tilde{p}_{t_1}, \tilde{p}_{t_2} | \mathcal{I}_{t_0}]$, each quantity being separately obtainable from the Gaussian process forecast, *cf.* eq. (6). The trade decision is made in one of two ways, depending on whether a position has already been opened:

1. When making a decision at time t_0, if a position has **not yet been entered** for the spread in a given trading year, eq. (26) is maximized with respect to unconstrained $t_1, t_2 \geq t_0$. An illustration of this criterion is given in Figure 9, which corresponds to the first decision made when trading the spread shown in Figure 6.
2. In contrast, if a position **has already been opened**, eq. (26) is only maximized with respect to t_2, keeping t_1 fixed at t_0. This corresponds to revising the exit point of an existing position.

[16] An *information ratio* is defined as the average return of a portfolio in excess of a benchmark, divided by the standard deviation of the excess return distribution; see (Grinold and Kahn, 2000) for more details.

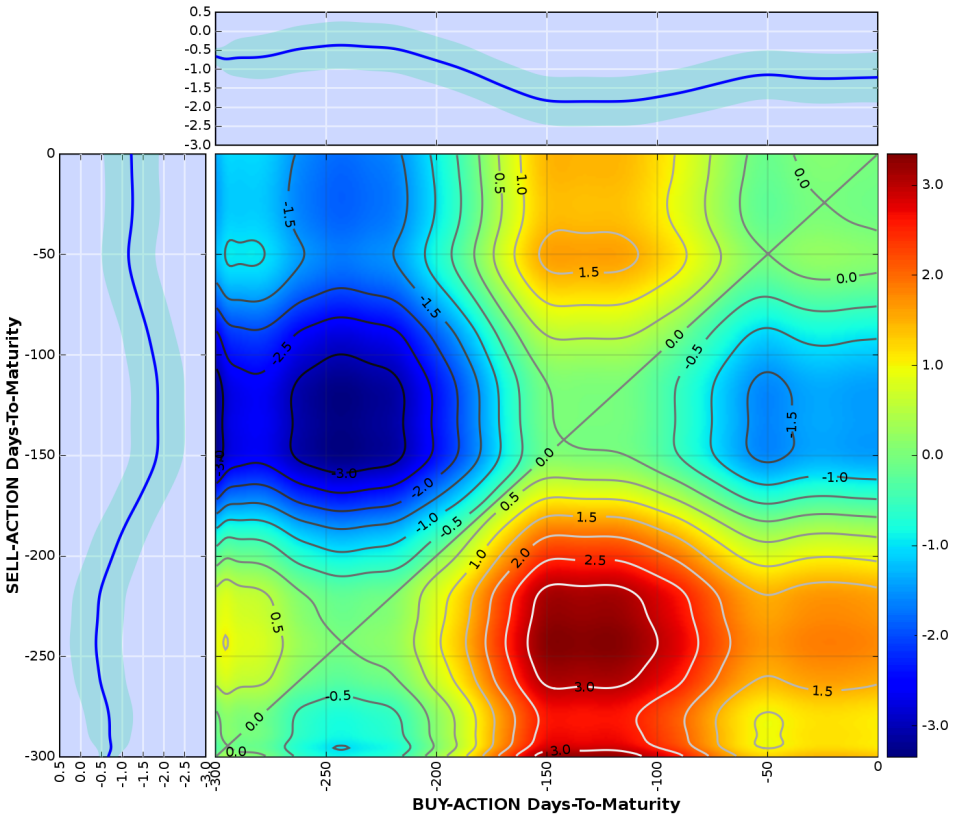

Figure 9: Computation of the Information Ratio between each potential entry and exit points, given the spread trajectory forecast. The x-axis represents the potential "buy" times, and the y-axis the potential "sell" times.

Simple additional filters are used to avoid entering marginal trades: we impose a trade duration of at least four days, a minimum forecast IR of 0.25 and a forecast standard deviation of the price sequence of at least 0.075. These thresholds have not been tuned extensively; they were used only to avoid trading on an approximately flat price forecast.

We observe in passing that this framework is vastly different from a classical mean-variance portfolio optimization (Markowitz, 1959), which would

Table 2: List of spreads used to construct the tested portfolio.

Commodity	**Maturities *short–long***
Cotton	10–12, 10–5
FeederCattle	11–3, 8–10
Gasoline	1–5
LeanHogs	12–2, 2–4, 2–6, 4–6, 7–12, 8–10, 8–12
LiveCattle	2–8
NaturalGas	6–9, 6–12
SoybeanMeal	5–9, 7–9, 7–12, 8–3, 8–12
Soybeans	1–7, 7–11, 7–9, 8–3, 8–11
Wheat	3–7, 3–9, 5–9, 5–12, 7–12

focus on *single-period* expected returns and variance thereof. The proposed framework for trading spreads, by relying on an explicit forecast of a *complete future trajectory*, is intuitive to practitioners, who can readily pass judgement about the accuracy of a forecast and the resulting suggested trades based on their own market view.

8. Financial Performance

We applied these ideas to trading an equally-weighted portfolio of 30 spreads, selected among the following commodities: Cotton (2 spreads), Feeder Cattle (2), Gasoline (1), Lean Hogs (7), Live Cattle (1), Natural Gas (2), Soybean Meal (5), Soybeans (5), Wheat (5); the complete list of spreads traded appears in Table 2. The spreads were selected on the basis of their performance on the 1994–2002 period.

Our simulations were carried on the 1994–2007 period, using historical

data (for Gaussian process training) dating back to 1989. Transaction costs were assumed to be 5 basis points per spread leg traded. Spreads were never traded later than 25 calendar days before maturity of the near leg. Relative returns are computed using as a notional amount half the total exposure incurred by both legs of the spread.[17] Moreover, since individual spreads trade only for a fraction of the year, returns are taken into consideration for a spread only when the spread is actually trading.

Financial performance results on the complete test period and two disjoint sub-periods (which correspond, until end-2002 to the model selection period, and after 2003 to a true out-of-sample evaluation) are shown in Tables 3 to 5. All results give returns in excess of the risk-free rate (since margin deposits earn interest in a futures trading account). In all sub-periods, but particularly since 2003, the portfolio exhibits a very favorable risk-return profile, including positive skewness and acceptable excess kurtosis.[18]

Year-by-year returns, for the complete portfolio as well as sub-portfolios formed by the spreads on a single underlying commodity, are given in Table 6. Furthermore, the monthly-return correlation matrix among the sub-portfolios formed by same-commodity spreads appears is Table 7.

A plot of cumulative returns, number of open positions and monthly returns appears in Figure 10.

[17] This is a conservative assumption, since most exchanges impose considerably reduced margin requirements on recognized spreads.

[18] By way of comparison, over the period 1 Jan. 1994–30 Apr. 2007, the S&P 500 index has an information ratio of approximately 0.37 against the U.S. three-month treasury bills.

Table 3: Performance on the 1 Jan. 1994–30 April 2007 period. All returns are expressed in excess of the risk-free rate. The information ratio statistics are annualized. Skewness and excess kurtosis are on the monthly return distributions. Drawdown duration is expressed in calendar days. The model shows good performance for moderate risk.

	Portfolio	Cotton	F. Cattle	Gasoline	Lean Hogs
Annualized Return	7.4%	1.6%	1.8%	2.5%	5.2%
Avg Annual Return	7.3%	3.3%	2.6%	6.7%	5.6%
Avg Annual Stddev	4.1%	4.8%	4.0%	6.8%	8.3%
Annual Inf Ratio	1.77	0.68	0.65	0.99	0.67
Avg Monthly Return	0.6%	0.3%	0.2%	0.6%	0.5%
Avg Monthly Stddev	1.2%	1.4%	1.1%	2.0%	2.4%
Skewness	0.68	0.47	−0.07	0.13	0.32
Excess Kurtosis	3.40	2.63	4.04	0.66	1.11
Best Month	6.0%	5.4%	4.0%	6.1%	9.3%
Worst Month	−3.4%	−4.0%	−4.9%	−4.7%	−5.9%
Fraction Months Up	71%	56%	50%	67%	58%
Max. Drawdown	−7.7%	−6.8%	−7.0%	−6.5%	−12.9%
Drawdown Duration	653	705	774	1040	137
Drawdown From	1997/02	1997/05	1994/08	1996/11	1998/09
Drawdown Until	1998/11	1999/04	1996/09	1999/09	1999/01
	L. Cattle	Nat. Gas	SB Meal	Soybeans	Wheat
Annualized Return	1.5%	2.4%	2.8%	3.3%	7.0%
Avg Annual Return	3.8%	4.7%	4.3%	4.1%	8.6%
Avg Annual Stddev	7.6%	7.8%	8.8%	6.6%	6.5%
Annual Inf Ratio	0.50	0.60	0.49	0.61	1.33
Avg Monthly Return	0.3%	0.4%	0.4%	0.3%	0.7%
Avg Monthly Stddev	2.2%	2.2%	2.5%	1.9%	1.9%
Skewness	−0.35	0.74	2.23	−0.14	1.02
Excess Kurtosis	2.50	2.01	11.73	5.39	5.31
Best Month	5.7%	9.4%	14.1%	8.5%	9.5%
Worst Month	−8.1%	−5.2%	−6.9%	−7.7%	−6.5%
Fraction Months Up	58%	53%	53%	56%	71%
Max. Drawdown	−13.3%	−14.4%	−25.1%	−16.3%	−13.7%
Drawdown Duration	> 1286	1528	2346	1714	82
Drawdown From	2003/10	2001/02	1996/11	1996/12	2002/10
Drawdown Until	None	2005/04	2003/05	2001/09	2002/12

Table 4: Performance on the 1 January 1994–31 December 2002 period. All returns are expressed in excess of the risk-free rate. The information ratio statistics are annualized. Skewness and excess kurtosis are on the monthly return distributions. Drawdown duration is expressed in calendar days.

	Portfolio	Cotton	F. Cattle	Gasoline	Lean Hogs
Annualized Return	5.9%	2.0%	0.7%	1.6%	5.0%
Avg Annual Return	5.9%	4.3%	1.1%	4.5%	5.4%
Avg Annual Stddev	4.0%	5.3%	4.1%	7.0%	8.5%
Annual Inf Ratio	1.45	0.80	0.27	0.65	0.64
Avg Monthly Return	0.5%	0.4%	0.1%	0.4%	0.5%
Avg Monthly Stddev	1.2%	1.5%	1.2%	2.0%	2.4%
Skewness	0.65	0.42	−0.35	0.10	0.59
Excess Kurtosis	4.60	2.20	4.83	0.77	1.60
Best Month	6.0%	5.4%	4.0%	6.1%	9.3%
Worst Month	−3.4%	−4.0%	−4.9%	−4.7%	−5.4%
Fraction Months Up	67%	59%	47%	65%	58%
Max. Drawdown	−7.7%	−6.8%	−7.0%	−6.5%	−12.9%
Drawdown Duration	653	705	774	1040	137
Drawdown From	1997/02	1997/05	1994/08	1996/11	1998/09
Drawdown Until	1998/11	1999/04	1996/09	1999/09	1999/01

	L. Cattle	Nat. Gas	SB Meal	Soybeans	Wheat
Annualized Return	2.7%	0.5%	−0,2%	2.3%	5.6%
Avg Annual Return	6.2%	1.2%	−0,1%	2.8%	7.2%
Avg Annual Stddev	6.6%	8.2%	6,5%	5.5%	6.5%
Annual Inf Ratio	0.94	0.14	−0,01	0.52	1.11
Avg Monthly Return	0.5%	0.1%	0,0%	0.2%	0.6%
Avg Monthly Stddev	1.9%	2.4%	1,9%	1.6%	1.9%
Skewness	0.64	1.13	−0,70	−1.19	0.91
Excess Kurtosis	0.41	3.02	3,38	6.97	6.93
Best Month	5.7%	9.4%	5,7%	5.2%	9.5%
Worst Month	−2.6%	−5.2%	−6,9%	−7.7%	−6.5%
Fraction Months Up	62%	46%	51%	57%	68%
Maximum Drawdown	−8.2%	−11.7%	−25,1%	−16.3%	−13.7%
Drawdown Duration	485	> 686	> 2225	1714	82
Drawdown From	1997/08	2001/02	1996/11	1996/12	2002/10
Drawdown Until	1998/12	None	None	2001/09	2002/12

Table 5: Performance on the 1 January 2003–30 April 2007 period. All returns are expressed in excess of the risk-free rate. The information ratio statistics are annualized. Skewness and excess kurtosis are on the monthly return distributions. Drawdown duration is expressed in calendar days.

	Portfolio	Cotton	F. Cattle	Gasoline	Lean Hogs
Annualized Return	10.5%	0.6%	4.1%	4.4%	5.8%
Avg Annual Return	10.1%	1.2%	5.4%	10.9%	6.0%
Avg Annual Stddev	4.1%	3.6%	3.6%	6.3%	8.1%
Annual Inf Ratio	2.44	0.34	1.49	1.73	0.73
Avg Monthly Return	0.8%	0.1%	0.4%	0.9%	0.5%
Avg Monthly Stddev	1.2%	1.0%	1.0%	1.8%	2.3%
Skewness	0.76	0.11	0.89	0.35	−0.28
Excess Kurtosis	1.26	0.07	0.24	−0.30	−0.15
Best Month	4.8%	2.7%	3.3%	5.1%	5.0%
Worst Month	−1.8%	−2.0%	−1.3%	−2.5%	−5.9%
Fraction Months Up	77%	50%	56%	71%	58%
Maximum Drawdown	−4.0%	−5.8%	−4.1%	−4.4%	−11.1%
Drawdown Duration	23	309	373	297	355
Drawdown From	2004/06	2003/08	2004/08	2003/09	2004/03
Drawdown Until	2004/07	2004/06	2005/08	2004/07	2005/03
	L. Cattle	Nat. Gas	SB Meal	Soybeans	Wheat
Annualized Return	−0.9%	6.3%	9.2%	5.5%	9.8%
Avg Annual Return	−1.8%	11.0%	12.0%	6.4%	11.3%
Avg Annual Stddev	9.5%	6.8%	11.6%	8.4%	6.4%
Annual Inf Ratio	−0.19	1.62	1.04	0.76	1.77
Avg Monthly Return	−0.2%	0.9%	1.0%	0.5%	0.9%
Avg Monthly Stddev	2.7%	2.0%	3.3%	2.4%	1.8%
Skewness	−0.86	−0.09	2.55	0.30	1.25
Excess Kurtosis	1.66	−0.10	7.18	2.73	1.67
Best Month	5.3%	4.7%	14.1%	8.5%	6.6%
Worst Month	−8.1%	−3.6%	−3.9%	−7.3%	−2.1%
Fraction Months Up	50%	66%	58%	55%	77%
Maximum Drawdown	−13.3%	−6.9%	−13.4%	−13.0%	−9.3%
Drawdown Duration	> 1286	112	16	334	69
Drawdown From	2003/10	2003/01	2004/06	2003/04	2003/10
Drawdown Until	None	2003/04	2004/07	2004/03	2003/12

Table 6: Yearly returns of the entire portfolio and of sub-portfolios formed by spreads on the same underlying commodity. Year 2007 includes data until April 30 (the return reported is not annualized).

	Portfolio	Cotton	F.Cattle	Gasoline	LeanHogs	L.Cattle	N.Gas	SB Meal	Soybeans	Wheat
1994	3.9%	0.0%	0.1%	0.0%	−0.6%	7.7%	−0.2%	−0.6%	−1.3%	8.6%
1995	4.0%	5.8%	1.5%	0.0%	4.0%	2.9%	−4.5%	1.9%	−1.4%	−0.8%
1996	5.0%	3.6%	−2.2%	−1.5%	−2.3%	1.5%	3.5%	3.8%	8.1%	0.6%
1997	−4.3%	0.1%	0.5%	2.8%	3.5%	−1.0%	−1.6%	−19.6%	−12.3%	14.5%
1998	7.1%	1.2%	2.4%	−0.5%	4.2%	5.5%	6.8%	−4.5%	2.7%	5.5%
1999	10.1%	0.6%	2.0%	4.6%	16.1%	2.5%	7.1%	4.7%	0.0%	4.1%
2000	1.9%	1.5%	0.4%	−1.4%	−1.6%	−4.8%	−1.2%	9.1%	4.7%	0.7%
2001	8.7%	2.7%	2.2%	10.8%	8.8%	3.3%	−5.0%	−0.9%	8.5%	7.0%
2002	16.2%	2.5%	−0.8%	0.0%	13.1%	6.4%	0.2%	7.7%	13.2%	9.9%
2003	10.8%	2.5%	8.2%	1.7%	4.4%	−2.8%	1.0%	3.4%	−1.4%	18.2%
2004	14.7%	−1.5%	−0.8%	7.8%	0.0%	0.7%	3.5%	40.7%	24.5%	4.1%
2005	8.5%	3.2%	5.0%	6.9%	7.8%	1.8%	5.4%	5.4%	2.5%	4.4%
2006	10.3%	−1.8%	3.6%	2.8%	13.0%	−1.1%	17.7%	−4.1%	0.0%	12.8%
2007	1.4%	0.3%	1.8%	0.0%	0.3%	−2.3%	0.5%	−0.8%	0.1%	3.5%

Table 7: Correlation matrix of monthly returns among sub-portfolios formed by spreads on the same underlying commodity, on the 1994–2007 period.

	Cotton	F.Cattle	Gasoline	LeanHogs	L.Cattle	N.Gas	SB Meal	Soybeans	Wheat
Cotton	—	0.06	−0.11	−0.04	−0.02	0.02	0.11	0.07	−0.09
FeederCattle	0.06	—	0.12	−0.05	−0.04	−0.05	−0.02	−0.09	0.00
Gasoline	−0.11	0.12	—	0.01	0.21	−0.06	0.04	−0.04	−0.02
LeanHogs	−0.04	−0.05	0.01	—	−0.14	−0.02	0.00	−0.02	0.01
LiveCattle	−0.02	−0.04	0.21	−0.14	—	−0.03	−0.01	−0.03	0.08
NaturalGas	0.02	−0.05	−0.06	−0.02	−0.03	—	−0.02	0.07	−0.13
SoybeanMeal	0.11	−0.02	0.04	0.00	−0.01	−0.02	—	0.40	−0.02
Soybeans	0.07	−0.09	−0.04	−0.02	−0.03	0.07	0.40	—	−0.04
Wheat	−0.09	0.00	−0.02	0.01	0.08	−0.13	−0.02	−0.04	—

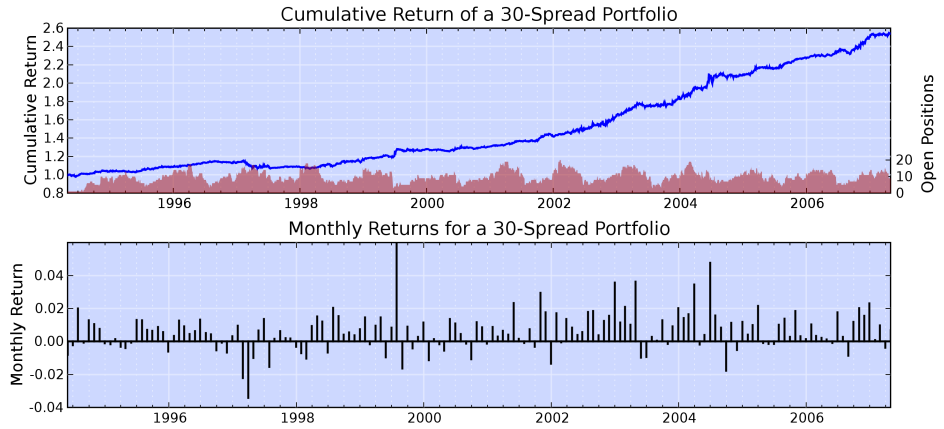

Figure 10: **Top Panel:** cumulative excess return after transaction costs of a portfolio of 30 spreads traded according to the maximum information-ratio criterion; the bottom part plots the number of positions open at a time (right axis). **Bottom Panel:** monthly portfolio relative excess returns; we observe the significant positive skewness in the distribution.

9. Conclusions

We introduced a flexible functional representation of time series, capable of making long-term forecasts from progressively-revealed information sets and of handling multiple irregularly-sampled series as training examples. We demonstrated the approach on a challenging commodity spread trading application, making use of a Gaussian process' ability to compute a complete covariance matrix between several test outputs.

Future work includes making more systematic use of approximation methods for Gaussian processes (see Quiñonero-Candela and Rasmussen (2005) for a survey). The specific usage pattern of the Gaussian process may guide the approximation: in particular, since we know in advance the test inputs, the problem is intrinsically one of *transduction* (Vapnik, 1998), and the Bayesian Committee Machine (Tresp, 2000) could prove beneficial.

10. Appendix: Analysis of Forecasting Errors for Wheat 3–7

The following pages illustrate with more detail the forecasting performance of the models compared in §6/p. 35, for every year of the period 1994–2007 (until Apr. 30, for 2007), for the March–July Wheat spread.

We focus on the results (both for the standardised squared error and standardised negative log likelihood) *as a function of the forecast horizon* (in calendar days). Figures 11 and 12 give absolute performance figures for all models being compared and all years. The solid red red line is the result of a local smoothing of the error and is useful to identify trends.

As a general rule, we note that performance degrades as the forecast horizon increases, as can be inferred from the upward-sloping trend line. Also, the augmented-representation Gaussian process with all variables, **AugRQ/all-inp** (the "reference model"), appears to be systematically as good or better than the other models.

This intuition is confirmed by Figures 13 to **??**, which compare, in a pairwise fashion, **AugRQ/all-inp** against every other model. Performance measures (and the smoothed trend line) below the zero line indicate that **AugRQ/all-inp** performs better than the alternative.

In many contexts, we see that this model beats the alternative (trend line below zero), but — just as significantly — its advantage increases with the forecast horizon, as witnessed by the *downward-sloping* trend line. In other words, even though the performance of both models generally tends to decrease with the horizon, **AugRQ/all-inp** generally holds its own better against the alternative and degrades less rapidly.

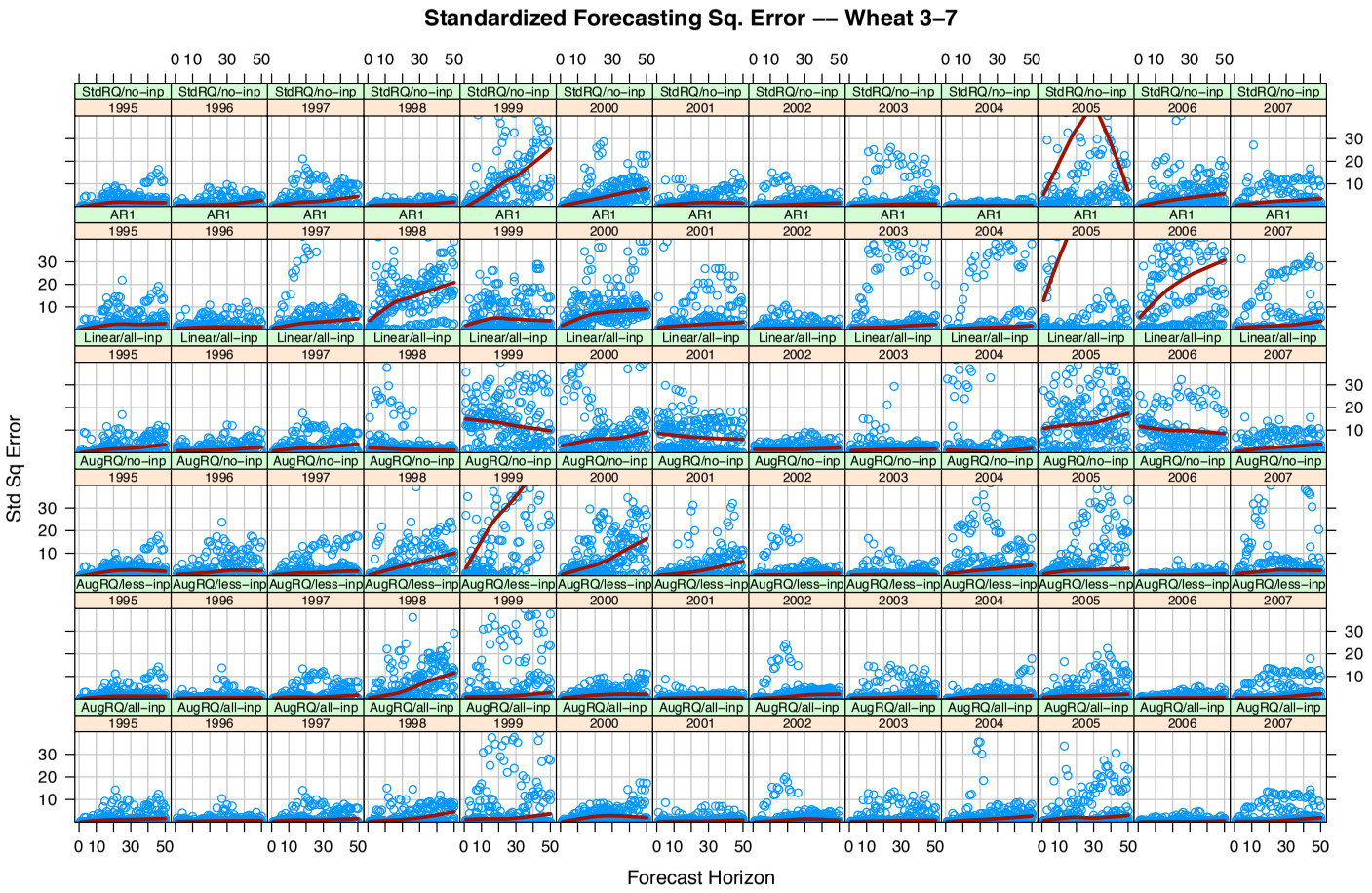

Figure 11: Standardized forecasting squared error as a function of the forecast horizon (calendar days) on the March–July Wheat spread, for all models considered in the main text and all years of the test period. We note that the mean forecast error tends to increase with the horizon.

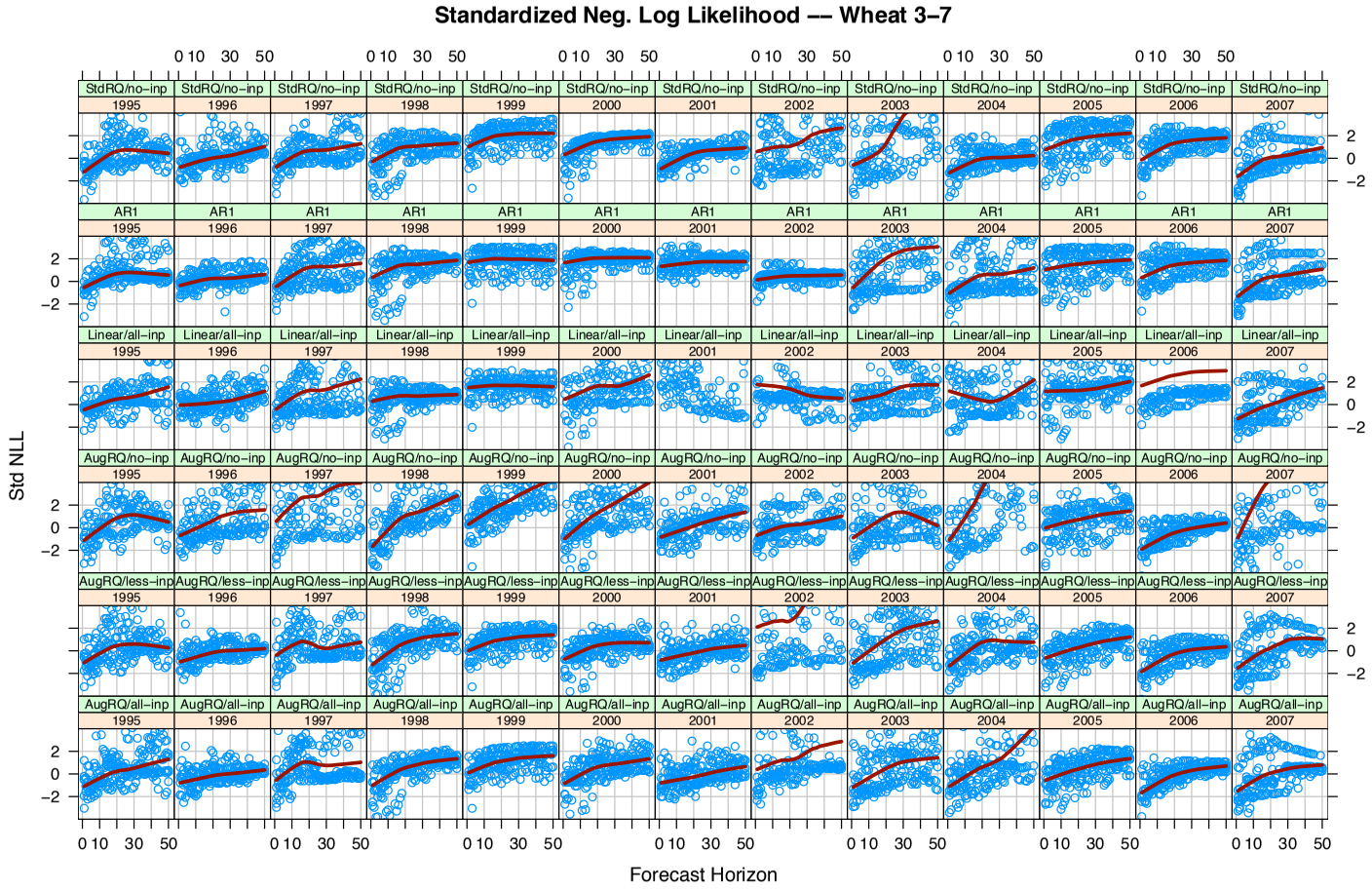

Figure 12: Standardized forecasting negative log likelihood as a function of the forecast horizon (calendar days) on the March–July Wheat spread, for all models considered in the main text and all years of the test period. We note that the mean forecast error tends to increase with the horizon.

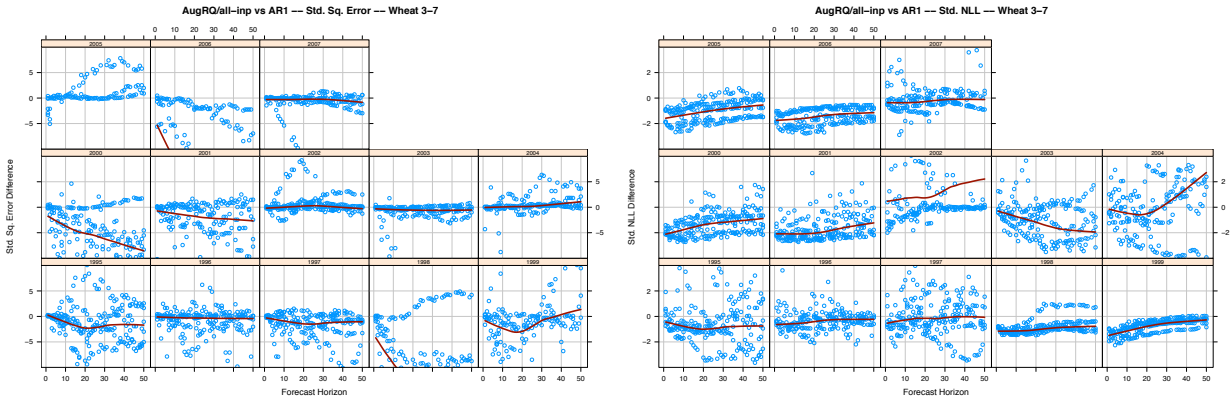

Figure 13: Squared error difference (left) and NLL difference (right) between **AugRQ/all-inp** and **AR1** as a function of the forecast horizon, for each year of the test period. A negative difference indicates an advantage for **AugRQ/all-inp**.

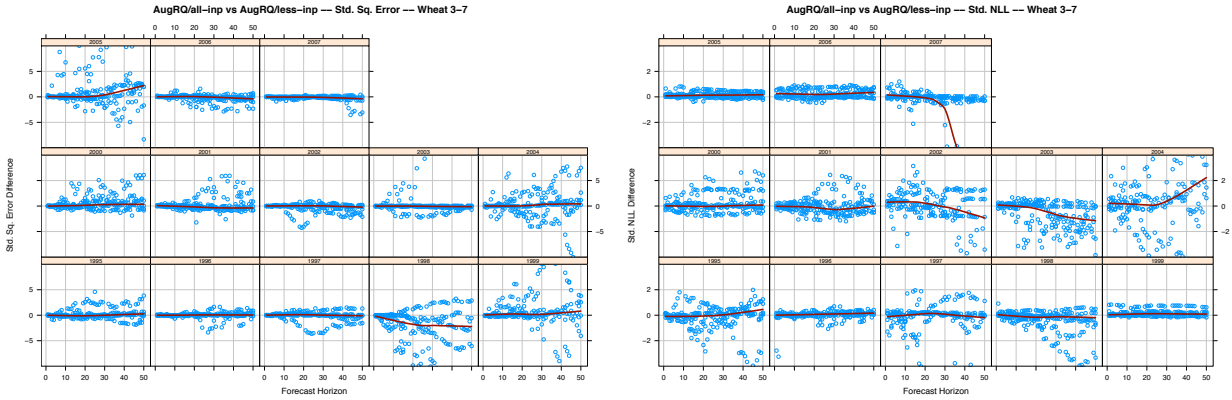

Figure 14: Squared error difference (left) and NLL difference (right) between **AugRQ/all-inp** and **AugRQ/less-inp** as a function of the forecast horizon, for each year of the test period. A negative difference indicates an advantage for **AugRQ/all-inp**.

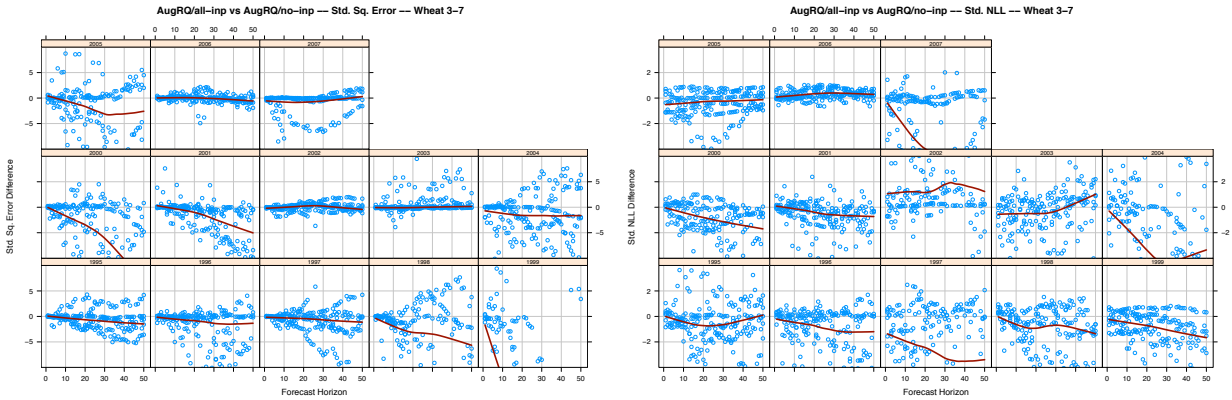

Figure 15: Squared error difference (left) and NLL difference (right) between **AugRQ/all-inp** and **AugRQ/no-inp** as a function of the forecast horizon, for each year of the test period. A negative difference indicates an advantage for **AugRQ/all-inp**.

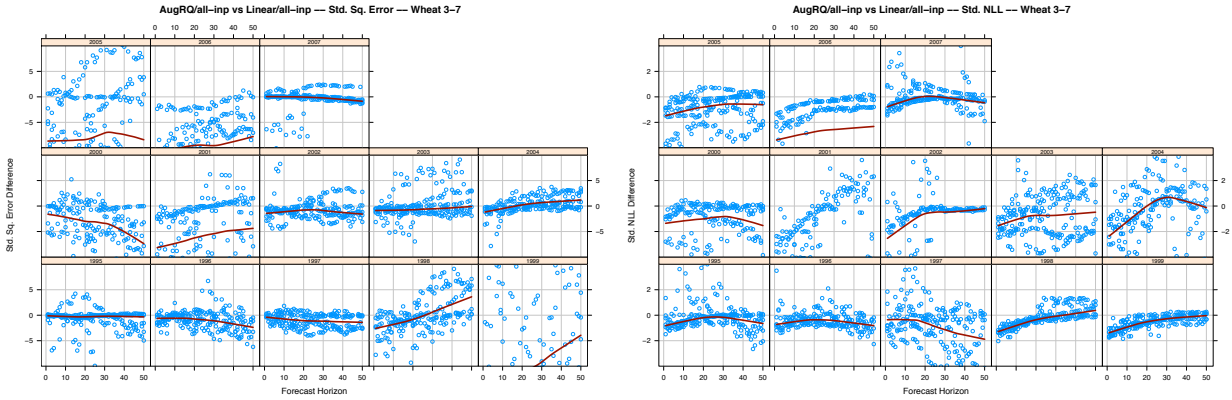

Figure 16: Squared error difference (left) and NLL difference (right) between **AugRQ/all-inp** and **Linear/all-inp** as a function of the forecast horizon, for each year of the test period. A negative difference indicates an advantage for **AugRQ/all-inp**.

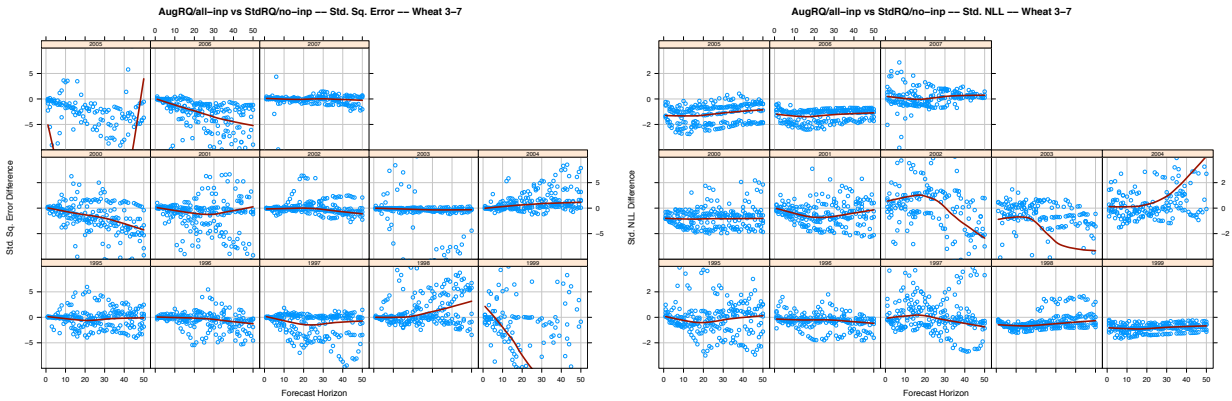

Figure 17: Squared error difference (left) and NLL difference (right) between **AugRQ/all-inp** and **StdRQ/no-inp** as a function of the forecast horizon, for each year of the test period. A negative difference indicates an advantage for **AugRQ/all-inp**.

References

Bertsekas, D. P., 2000. Nonlinear Programming, 2nd Edition. Athena Scientific, Belmont, MA.

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Oxford University Press.

Butterworth, D., Holmes, P., 2002. Inter-market spread trading: evidence from uk index futures markets. Applied Financial Economics 12 (11), 783–790.

Chapados, N., Bengio, Y., July 2001. Cost functions and model combination for VaR-based asset allocation using neural networks. IEEE Transactions on Neural Networks 12 (4), 890–906.

Clements, M. P., Hendry, D. F., 1998. Forecasting Economic Time Series. The Marshall Lectures on Economic Forecasting. Cambridge University Press, Cambridge, UK.

Cressie, N. A. C., 1993. Statistics for Spatial Data. Wiley.

Diebold, F. X., Mariano, R. S., July 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13 (3), 253–263.

Dunis, C. L., Laws, J., Evans, B., January 2006a. Trading futures spread portfolios: Applications of higher order and recurrent networks. Tech. rep., Centre for International Banking, Economics and Finance; Liverpool John Moores University, `www.cibef.com`.

Dunis, C. L., Laws, J., Evans, B., 2006b. Trading futures spreads: an appli-

cation of correlation and threshold filters. Applied Financial Economics 16 (12), 903–914.

Dutt, H. R., Fenton, J., Smith, J. D., Wang, G. H., 1997. Crop year influences and variability of the agricultural futures spreads. The Journal of Futures Markets 17 (3), 341–367.

Girard, A., Rasmussen, C. E., Candela, J. Q., Murray-Smith, R., 2003. Gaussian process priors with uncertain inputs – application to multiple-step ahead time series forecasting. In: S. Becker, S. T., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems 15. MIT Press, pp. 529–536.

Girma, P. B., Paulson, A. S., 1998. Seasonality in petroleum futures spreads. The Journal of Futures Markets 18 (5), 581–598.

Grinold, R. C., Kahn, R. N., 2000. Active Portfolio Management. McGraw Hill.

Hamilton, J. D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Hull, J. C., 2005. Options, Futures and Other Derivatives, sixth Edition. Prentice Hall, Englewood Cliffs, NJ.

Kim, M. K., Leuthold, R. M., April 2000. The distributional behavior of futures price spreads. Journal of Agricultural and Applied Economics 32 (1), 73–87.

Liu, S.-M., Chou, C.-H., 2003. Parities and spread trading in gold and silver markets: A fractional cointegration analysis. Applied Financial Economics 13 (12), 879–891.

MacKay, D. J. C., 1994. Bayesian methods for backprop networks. In: Domany, E., van Hemmen, J. L., Schulten, K. (Eds.), Models of Neural Networks, III. Springer, pp. 211–254.

MacKay, D. J. C., 1999. Comparison of approximate methods for handling hyperparameters. Neural Computation 11 (5), 1035–1068.

Markowitz, H. M., 1959. Portfolio Selection: Efficient Diversification of Investment. John Wiley & Sons, New York, London, Sydney.

Matheron, G., 1973. The intrinsic random functions and their applications. Advances in Applied Probability 5, 439–468.

Neal, R. M., 1996. Bayesian Learning for Neural Networks. Lecture Notes in Statistics 118. Springer.

O'Hagan, A., 1978. Curve fitting and optimal design for prediction. Journal of the Royal Statistical Society B 40, 1–42, (With discussion).

Poitras, G., 1987. "golden turtle tracks": In search of unexploited profits in gold spreads. The Journal of Futures Markets 7 (4), 397–412.

Quiñonero-Candela, J., Rasmussen, C. E., 2005. A unifying view of sparse approximate gaussian process regression. Journal of Machine Learning Research 6, 1939–1959.

Ramsay, J. O., Silverman, B. W., 2005. Functional Data Analysis, 2nd Edition. Springer.

Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. MIT Press.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Simon, D. P., 1999. The soybean crush spread: Empirical evidence and trading strategies. The Journal of Futures Markets 19 (3), 271–389.

Stock, J. H., Watson, M. W., 1999. Forecasting inflation. Journal of Monetary Economics 44, 293–335.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society, B 36 (1), 111–147.

Sundararajan, S., Keerthi, S. S., 2001. Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. Neural Computation 13 (5), 1103–1118.

Tresp, V., 2000. A bayesian committee machine. Neural Computation 12, 2719–2741.

U.S. Department of Agriculture, 2008. Economic research service data sets. WWW publication, available at `http://www.ers.usda.gov/Data/`.

Vapnik, V., 1998. Statistical Learning Theory. John Wiley & Sons, New York, London, Sydney.

Wahab, M., Cohn, R., Lashgari, M., 1994. The gold-silver spread: Integration, cointegration, predictability, and ex-ante arbitrage. The Journal of Futures Markets 14 (6), 709–756.

Williams, C. K. I., Rasmussen, C. E., 1996. Gaussian processes for regression. In: Touretzky, D. S., Mozer, M. C., Hasselmo, M. E. (Eds.), Ad-

vances in Neural Information Processing Systems 8. MIT Press, pp. 514–520.

Working, H., 1934. Price relationships between may and new-crop wheat future at chicago since 1885. Wheat Studies 10, 183–228.

Working, H., oct 1935. Differential price behavior as a subject for commodity price analysis. Econometrica 3 (4), 416–427.

Working, H., dec 1949. The theory of price of storage. The American Economic Review 39 (6), 1254–1262.