Math Finan Econ (2018) 12:97–109
https://doi.org/10.1007/s11579-017-0198-8

Combining different models

L. C. G. Rogers[1]

Received: 29 August 2016 / Accepted: 9 August 2017 / Published online: 18 August 2017

Abstract Portfolio selection is one of the most important areas of modern finance, both theoretically and practically. Reliance on a single model is fraught with difficulties, so attempting to combine the strengths of different models is attractive; see, for example, Geweke and Amisano (J Econom 164(1):130–141, 2011) and the many references therein. This paper contributes to the model combination literature, but with a difference: the models we consider here are making statements about *different* sets of assets. There appear to be no studies making this structural assumption, which completely changes the nature of the problem. This paper offers suggestions for principles of model combination in this situation, characterizes the solution in the case of multivariate Gaussian distributions, and provides a small illustrative example.

Keywords Model · Combination · Bayes · Entropy minimization · Marginal

JEL Classification C02 · C18 · C38 · C55 · G11

1 Introduction

Suppose that you are faced with the problem of choosing a portfolio position in a universe of N assets, where N may be many hundreds. It is generally understood that a simple-minded direct attempt to build a portfolio involving all N of the assets will be a dismal failure, for various reasons, chief among them being the difficulty in forming accurate estimates of the covariance matrix of returns; see, for example, the book by Fan et al. [2]. The dimensionality of the problem requires innovation, and there are many different directions we may look to get traction. For example, we might propose a low-dimensional factor model, where all the returns processes are driven by a handful of factors which should be easier to deal with. The

✉ L. C. G. Rogers
l.c.g.rogers@statslab.cam.ac.uk

[1] University of Cambridge, Cambridge, UK

factors could be series which are economically significant, such as the returns on a major stock index, the prices of important commodities, key interest rates or exchange rates; or the factors could be derived from a principal components analysis of an estimated covariance matrix. Or again, we might constrain the portfolio positions to be long-only to try to deal with the extreme long-short positions that generally arise in a simple-minded approach. We might increment the covariance with a multiple of the identity, for the same reason. We might take different models and combine their forecasts in some way, as in Bates and Granger [3], Elliott and Timmermann [4], Geweke and Amisano [1], Pettenuzzo and Ravazzolo [5], and many other papers referred to therein.

A different approach to the high-dimensionality would be to split the universe of assets into smaller sets of assets, and try to do something sensible with those smaller sets. If we believe we can make a reasonable combination of up to ten assets (say) then we could in principle use such a 'divide and conquer' approach, but it would not allow us to exploit the correlations *between* sets of assets, and the problem still remains of how to weight the different portfolios formed from the subsets.

The approach taken here has this flavour, in that we suppose the universe of N assets is broken down into subsets of assets, but we do not suppose that those subsets are *disjoint.*[1] It might be for example that we want to build one model for G10 currencies, and another one for European government bonds, stock indices and currencies; we may have insights into the way currencies move together, and we may have separate insights into how a nation's currency, stock index and bonds move together. Now we would like to combine these (hopefully reasonably good) models. So the European G10 currencies are common to both, but each model speaks of variables that are outside the other. How should this be done?

This paper offers some possible answers. In Sect. 2, we introduce notation and formulate the problem. Models speak of different sets of assets, and this introduces an equivalence relation on the assets, two assets being considered equivalent if there is no model which speaks about one of the assets but not the other. The equivalence classes (which for brevity we refer to as tiles) are sets of assets that can be considered together for the purposes of inference. We then propose that all those models which speak about the assets in a given tile are combined by Bayesian model averaging[2]. This tells us how to combine the statements of all our models for any individual tile, but how we combine across different tiles still needs to be specified. We address this in Sect. 3, where we propose to construct a measure with the required tile marginals which solves a *relative entropy minimization problem*. It turns out that this problem has a unique solution, which can be characterized quite cleanly, albeit implicitly.

This approach operates at an abstract level, so we do not need to make any structural assumptions about any of the variables; we do not even need to assume that they take values in a vector space. However, to be able to apply the results, we develop the form of the solution under the hypothesis that all the predictive distributions are *multivariate Gaussians*. The hypothesis is unlikely to hold in practical situations, but it is a plausible approximation, and we are able to make the combined distribution reasonably explicit. Identifying the distribution in general requires numerical solution. We briefly discuss a numerical example before concluding.

[1] Quite how this decomposition is to be done is not the subject of this paper; different finance industry professionals will have different views on what makes sense, and if I had my own unique insights into good ways of doing this decomposition, you may be sure that I would not be revealing them here. Suffice it to say that this study came from an industry context where exactly this question arose.

[2] This part of the story is hardly new; Geweke and Amisano [1] for example follow this route. We do propose a little variant (3) of the standard Bayesian approach which avoids the known issue that over time the posterior distribution converges to a point mass on just one model.

2 Problem formulation

We work in discrete time with a universe of N assets. The returns on day t will be denoted by the N-vector

$$X_t = \left(X_t^1, \ldots, X_t^N\right),$$

and we take the probability space to be the canonical path space $\Omega \equiv (\mathbb{R}^N)^{\mathbb{Z}^+}$. Information available by time t will be denoted by $\mathcal{F}_t$. Let the set $S = \{1, \ldots, N\}$ index the assets. Suppose that we have models $M^1, \ldots, M^K$, where model M^α makes predictions only about assets with labels in $S_\alpha \subseteq S$. To avoid triviality, we assume that $\cup_\alpha S_\alpha = S$. It is also worth noticing that if there is some set $I \subset \{1, \ldots, K\}$ of models such that $S_I \equiv \cup_{\alpha \in I} S_\alpha$ is disjoint from $S_{\sim I} \equiv \cup_{\alpha \notin I} S_\alpha$, then there is no connection between the models $\{M^\alpha,\ \alpha \in I\}$ and $\{M^\alpha,\ \alpha \notin I\}$. We could therefore analyse the two sets of models completely separately, and in practice it will probably be worth making such decompositions before we start, though in the account which follows this will not be assumed.

Formally, model M^α is a sequence $(m_t^\alpha)_{t=1,2,\ldots}$ where[3] for each $t \geq 1$

$$m_t^\alpha : \Omega \to \mathcal{P}(\mathbb{R}^{S_\alpha}) \textbf{ is } \mathcal{F}_{t-1} - \textbf{measurable}.$$

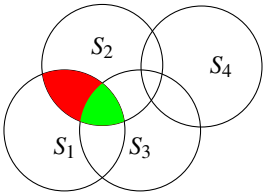
We refer to the m_t^α as the *predictive distributions* of the individual models; these are the central objects of study.

Notice that a model M^α is *not* a probability on $(\Omega, \mathcal{F})$, because it makes no statements about the distributions of returns X^i for $i \notin S_\alpha$. Nor is M^α a probability defined on the sub-sample space $\Omega_\alpha \equiv (\mathbb{R}^{S_\alpha})^{\mathbb{Z}^+}$, because it may be informed by the behaviour of assets other than those in S_α. The number N of assets may be very large, and each set S_α may be quite small, perhaps just a single asset. Sets S_α and $S_{\alpha'}$ may be disjoint, they may have non-empty intersection, one may be contained in the other; *anything is possible.*

On day $t-1$, each model M^α makes a prediction m_t^α of the distribution of the day-t returns of the assets in S_α; the goal is to find some algorithm to *combine the m_t^α into a single predictive distribution for all the assets in S* in a reasonable way. We shall require that the combination algorithm should not depend on specific distributional assumptions, and should be compatible with Bayesian principles. The set S is partitioned by the equivalence relation

$$i \sim j \Leftrightarrow \forall \alpha :\ \textbf{either} \text{ both } i,\ j \text{ are in } S_\alpha \textbf{ or } \text{ neither is in } S_\alpha. \tag{1}$$

into equivalence classes $C_k,\ k = 1, \ldots, J$, which we will refer to as *tiles*. These are sets of variables which are not split by any model. This simple Venn diagram illustrates the situation; the region coloured red is a tile, $C_r \equiv (S_1 \cap S_2) \backslash S_3$, as is the region coloured green, $C_g \equiv S_1 \cap S_2 \cap S_3$.



Thus on day $t-1$ each of the models $M^1,\ M^2,\ M^3$ makes a prediction for day t about the variables in the green tile: how should we combine these? Bearing in mind that we can

[3] $\mathcal{P}(Y)$ denotes the (Polish) space of all probability measures on (Polish) space Y.

only compare predictions which speak about the *same* variables, we propose the following principle for combining the predictive distributions:

(P0) Predictions for a tile are done by Bayesian model averaging over the maximal set of common variables.

So if we consider the green tile, models M^1, M^2, M^3 predict for those variables (and others besides), and what we do is take the *marginals* of the predictive distributions m_t^α ($\alpha = 1, 2, 3$) on the green tile C_g, and then do a Bayesian model averaging of these marginal distributions.

To explain this in a little more detail, if on day $t-1$ model α states that the law of observation X_t to be observed on day t will have density $f_t^\alpha(\cdot)$, then the posterior probabilities π_t^α of the different models update according to Bayes' rule:

$$\pi_t^\alpha \propto \pi_{t-1}^\alpha \, f_t^\alpha(X_t), \tag{2}$$

where the constant of proportionality is determined by the requirement that the π_t^α sum to 1. Thus on day $t-1$, each of the models M^1, M^2, M^3 states a (marginal) density for the variables[4] $X_t[S_\alpha]$, and the predicted law of these variables will be the average of those predicted densities, weighted according to the day-$(t-1)$ posterior. In practice, the updating relation (2) is modified to

$$\pi_t^\alpha \propto \sum_\beta \pi_{t-1}^\beta \, p_{\beta\alpha} \, f_t^\alpha(X_t), \tag{3}$$

where $P = (p_{jk})$ is a fixed transition matrix[5]. The interpretation is that the data-generating process may change state like a Markov chain with transition matrix P. The reason for introducing this possibility is because of the tendency of the updating recursion (2) to get stuck at historical average values, which in the context of asset returns is undesirable—we do not believe that asset returns from the distant past should have the same influence on our actions as more recent data, and using (3) instead of (2) reflects that.

To expand a little on this, the numerical example studied later performs a Bayesian averaging of models where the predicted mean return is a geometrically-weighted moving average of recent returns, with different weighting parameters for different models, perhaps with mean lookbacks of 10, 20 and 40 days. So each of these models pays more attention to recent returns than to older data, but what we want to avoid is the situation where after a lot of data has been processed we put almost all the weight on the model with a 20-day lookback, and are unable to change that posterior very much during the next few months. Experience shows that market dynamics can change quite quickly, and a model that does not admit that will go on losing money for too long when such a change happens.

This explains how we use the data gathered by day $t-1$ to state what we think the distribution of the C_g-variables will be on day t. One further point needs to be made, however, when we look at the variables in the red tile $C_r \equiv (S_1 \cap S_2)\backslash S_3$, because the procedure just detailed for tile C_g could be applied in two different ways: do we

- form the Bayesian average of M^1, M^2 on $S_1 \cap S_2$, and then take the $(S_1 \cap S_2)\backslash S_3$-marginal?
- **or** form the Bayesian average of M^1, M^2 on $(S_1 \cap S_2)\backslash S_3$?

Principle (P0) tells us to do the first of these; we form the Bayesian average of models M^1 and M^2 on *all* the variables they have in common, and then marginalize the distributional statement to those variables that they share with no other model.

[4] We use the (Python) convention that $X[A]$ denotes the subvector $\{X_i : i \in A\}$ of the vector X.

[5] When P is the identity, we recover (2).

To summarize then, we have just described how we construct the predictive distribution each day for the variables in *each of the tiles separately.* But as yet we have not determined how we combine these to make a predictive distribution for *all of the variables at once.* Doing this is clearly the essence of the problem, because we have to decide what the co-dependence of the variables in different tiles will be in order to make portfolio choices.

Remarks At first sight, we can easily reduce the problem to one where all of the predictive distributions speak about all of the variables—for any variables not in S_α, we just say that model M^α predicts a large-variance zero-mean return! This gets round the mathematical issues at the cost of turning the problem into nonsense. Why? Suppose that we have just two models, M^1 which makes statements about all of the asset returns, and M^2 which makes statements about only asset 1. We could expand M^2 to speak about all the assets by saying that its predictions for the others are just noise, but the problem is that if M^2 happens to predict asset 1 much better than M^1, then we would end up with most posterior weight on M^2. We would then believe that we had no useable information about any of the assets other than asset 1, whereas in fact M^1 might be telling us some quite valuable information about them.

3 Combining distributions

The situation then is this. We have for each model M^α a predictive distribution m_t^α for the S_α-variables on day t; we have for each tile C_j a predictive distribution q_t^j for the C_j-variables on day t, obtained by Bayesian model averaging. How do we come up with a distribution $q \in \mathcal{P}(\mathbb{R}^N)$ with the properties:

(P1) For each $j = 1, \ldots, J$, the C_j-marginal of q is q_t^j;
(P2) The co-dependence of the different tiles is inherited from the co-dependence of the m_t^α ?

For notational convenience, we will henceforth abbreviate m_t^α to m^α, and q_t^j to q^j, as the time index is immaterial. Of course, property (P2) of the construct q is not defined as yet, and the essential issue is to try to give a reasonable and precise meaning to this.

Let us first look only at the tiles $C_1, \ldots, C_J$ which make up S_α; we want to make a distribution μ for the variables $X[S_\alpha]$ which satisfies (P1) and (P2). We could of course ensure property (P1) simply by taking the product measure

$$\mu = q^1 \otimes \cdots \otimes q^J$$

but this ignores any information about co-dependence which there might be in m^α. We want to construct some measure μ which is as 'near' as possible to m^α while satisfying property (P1). The sense of closeness we propose here is closeness in *relative entropy*—other choices could be made, but this is a very natural one, and is widely used. So we will determine the measure μ by solving the problem

$$\min H(\mu|m^\alpha) \quad \text{subject to } \mu_{C_j} = q^j \quad \forall j, \tag{4}$$

where μ_A denotes the A-marginal of μ, and as usual

$$H(\mu|m) \equiv \int \frac{d\mu}{dm} \log\left(\frac{d\mu}{dm}\right) dm.$$

With a minor abuse of notation, we shall write

$$H(f|m) = \int f \log(f)\, dm \tag{5}$$

when f is a probability density with respect to m, or even $H(f)$ if the reference measure m does not need to be clarified.

To ensure that problem (4) is well-posed, we have the following result.

Proposition 1 *Assume that there exists some measure μ with $\mu_{C_j} = q^j$ for all $j = 1, \ldots, J$ for which the relative entropy $H(\mu|m^\alpha)$ is finite. Then the problem (4) has a unique solution.*

Proof Let $\mathcal{S}$ denote the set of all densities f for which the relative entropy

$$H(f|m^\alpha) \equiv \int f \log f\ dm^\alpha$$

is finite, and for which the C_j-marginal is the given measure q^j

$$\int f(x)\, m^\alpha(dx[\sim C_j]) = q^j(x[C_j]) \qquad \forall j, \tag{6}$$

where $x[\sim C_j]$ denotes the subvector of x on the complement of C_j. The set $\mathcal{S}$ is a convex subset of $L^1_+(m^\alpha)$, non-empty by assumption. If $b = \inf\{H(f) : f \in \mathcal{S}\}$, we can find $f_n \in \mathcal{S}$ such that $H(f_n) < b + 2^{-n}$. The family (f_n) is clearly uniformly integrable, so we may apply a version of Komlos' Theorem (see Lemma 2.1 in [6]) to conclude that there exist $g_n \in \text{conv}(f_n, f_{n+1}, \ldots)$ which converge in L^1 to some limit g. But since $\mathcal{S}$ is convex, the g_n are in $\mathcal{S}$, and it is immediate that the limit is also; the marginals are as given by (6). By Fatou's Lemma, we conclude that $H(g) = b$, and the infimum is attained.

Finally, if g_1 and g_2 are two minimizers, we have that for $\lambda \in [0, 1]$ the function $\bar{g} = \lambda g_1 + (1 - \lambda)g_2$ is in $\mathcal{S}$. Writing $\varphi(x) = x \log(x)$, we deduce that

$$b \le \int \varphi(\bar{g})\, dm^\alpha \le \int \{\lambda\varphi(g_1) + (1 - \lambda)\varphi(g_2)\}\, dm^\alpha = b. \tag{7}$$

Strict convexity of φ implies that $m^\alpha(g_1 = g_2) = 1$. □

This tells us what to do if we were just concerned with a single S^α which was split into tiles, but we need to deal with the whole set $S = \cup_\alpha S^\alpha$. What we propose to do therefore is to seek some probability μ so as to

$$\min \sum_{\alpha=1}^{K} H\left(\mu_{S^\alpha}|m^\alpha\right) \quad \text{subject to} \quad \mu_{C_j} = q^j \qquad \forall j. \tag{8}$$

The argument of Proposition 1 which showed that the minimum is attained by a unique minimizer runs into technical issues, which are illustrated by simple examples.

Example 1 Suppose that f_n is the density of a bivariate normal distribution with mean 0 and covariance

$$V = \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix},$$

where the ρ_n increase to 1. Then the 1-marginal densities of the f_n are a uniformly integrable family, as are the 2-marginal densities, but the family (f_n) is not uniformly integrable.

This shows that we will not be able to re-run the argument of Proposition 1 for problem (8), because this argument required uniform integrability of the densities, not just of their marginals.

A second issue arises, illustrated by the next example.

Example 2 Suppose that $K = 2$, $S^1 = \{1, 2\}$, $S^2 = \{2, 3\}$, and that the model distributions m^1, m^2 are both bivariate $N(0, I)$. Suppose further that the predictive distributions for each of the three tiles are a standard $N(0, 1)$ law. Then for any $a \in (-1, 1)$ the law

$$N\left(0, \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & 0 \\ a & 0 & 1 \end{pmatrix}\right) \tag{9}$$

is a minimizer of the objective (8) (it achieves value 0), and satisfies the marginal law constraints on each tile. Therefore we cannot hope for uniqueness of the solution to problem (8), and the family of solutions will not in general be uniformly integrable. We return to this example later once we have explained how we plan to get around these issues.

To deal with these issues then, we propose to modify the problem (8) to the following:

$$\min_{\mu} \sum_{\alpha=1}^{K} H\left(\mu_{S^\alpha} | m^\alpha\right) + \varepsilon H\left(\mu | \otimes_{j=1}^{J} q^j\right) \quad \text{subject to} \quad \mu_{C_j} = q^j \quad \forall j. \tag{10}$$

Here, $\varepsilon > 0$ is some chosen positive parameter. This choice is arbitrary, and undesirable; the inclusion of this term is required to prevent degeneracy. It may be that we can deal directly with problem (8), but Example 2 shows that even if we could prove that a minimizer exists, we would in general need to make an arbitrary choice of minimizer, so some arbitrary choice cannot be avoided—but, as we shall see, we may be able to get around this. Granted this modification of the objective, we have the following analogue of Proposition 1, whose proof follows the same lines as the proof of Proposition 1, and is therefore omitted. The key point is that the set of μ with finite relative entropy $H(\mu | \otimes_{j=1}^{J} q^j)$ is once again uniformly integrable.

Proposition 2 *Assume that there exists some measure μ with $\mu_{C_j} = q^j$ for all $j = 1, \ldots, J$ for which the relative entropy $H(\mu | \otimes_{j=1}^{J} q^j)$ is finite. Then the problem (10) has a unique solution.*

Can we get a clearer picture of what the solution to problem 10 looks like? We can indeed, but to describe it we need some notation. So suppose that with respect to some product measure dx the measures m^α have densities f_α, and the tile marginals q^j have densities φ_j. We shall abbreviate

$$\bar{\varphi}(x) \equiv \prod_j \varphi_j(x[C_j]) \tag{11}$$

for the reference density. Let g_α denote the S_α marginal density of μ, so that the objective to be minimized in (10) can be written

$$\begin{aligned} Z = &\sum_\alpha \int g_\alpha(x[S_\alpha]) \log\left(\frac{g_\alpha(x[S_\alpha])}{f_\alpha(x[S_\alpha])}\right) dx[S_\alpha] \\ &+ \varepsilon \int g(x)\{\log g(x) - \log \bar{\varphi}(x)\}\, dx, \end{aligned} \tag{12}$$

where of course the function g is the density of the unknown measure μ. The constraints

$$g_\alpha(x[S_\alpha]) = \int g(x)\, dx[\sim S_\alpha] \tag{13}$$

$$\varphi_j(x[C_j]) = \int g(x)\, dx[\sim C_j] \tag{14}$$

must also be satisfied. We then have the following characterization of the optimal density g.

Theorem 1 *The density g of the optimal solution to problem (10) is represented as*

$$g(x) \propto \exp\left[\frac{1}{1+\varepsilon} \sum_\alpha \log f_\alpha(x[S_\alpha]) + \frac{\varepsilon}{1+\varepsilon} \sum_j \{\, \varepsilon^{-1} \eta_j(x[C_j]) + \log \varphi_j(x[C_j]) \,\} \right], \tag{15}$$

where the functions η_j are determined up to additive constants by the constraints (14).

Proof We shall absorb the constraints (13) with Lagrange multipliers $\lambda_\alpha(x[S_\alpha])$ and the constraints (14) with Lagrange multipliers $\eta_j(x[C_j])$ to construct the Lagrangian

$$\begin{aligned}
L = Z &+ \sum_\alpha \int \lambda_\alpha(x[S_\alpha]) \left\{ g_\alpha(x[S_\alpha]) - \int g(x)\, dx[\sim S_\alpha] \right\} dx[S_\alpha] \\
&+ \sum_j \int \eta_j(x[C_j]) \left\{ \varphi_j(x[C_j]) - \int g(x)\, dx[\sim C_j] \right\} dx[C_j] \\
= &\sum_\alpha \int g_\alpha(x[S_\alpha]) \log\left(\frac{g_\alpha(x[S_\alpha])}{f_\alpha(x[S_\alpha])} \right) dx[S_\alpha] \\
&+ \varepsilon \int g(x) \{ \log g(x) - \log \bar{\varphi}(x) \}\, dx \\
&+ \sum_\alpha \int \lambda_\alpha(x[S_\alpha]) \left\{ g_\alpha(x[S_\alpha]) - \int g(x)\, dx[\sim S_\alpha] \right\} dx[S_\alpha] \\
&+ \sum_j \int \eta_j(x[C_j]) \left\{ \varphi_j(x[C_j]) - \int g(x)\, dx[\sim C_j] \right\} dx[C_j].
\end{aligned}$$

Minimizing L over g_α gives the first-order condition[6]

$$0 = 1 + \log(g_\alpha / f_\alpha) + \lambda_\alpha, \tag{16}$$

and minimizing L over g gives the first-order condition

$$0 = \varepsilon[\log g - \log \bar{\varphi} + 1] - \sum_\alpha \lambda_\alpha - \sum_j \eta_j. \tag{17}$$

From these conditions, we deduce expressions for g_α and g in terms of the unknown multipliers λ_α and η_j:

[6] …now omitting arguments except where the meaning is unclear…

$$g_\alpha(x[S_\alpha]) \propto f_\alpha(x[S_\alpha]) \exp\{-\lambda_\alpha(x[S_\alpha])\} \tag{18}$$

$$g(x) \propto \bar{\varphi}(x) \exp\left\{\varepsilon^{-1}\sum_\alpha \lambda_\alpha(x[S_\alpha]) + \varepsilon^{-1}\sum_j \eta_j(x[C_j])\right\}$$

$$= \exp\left\{\varepsilon^{-1}\sum_\alpha \lambda_\alpha(x[S_\alpha]) + \sum_j \left(\varepsilon^{-1}\eta_j + \log\varphi_j\right)(x[C_j])\right\}. \tag{19}$$

Integrating (19) over $x[\sim S_\alpha]$ reveals that

$$g_\alpha(x[S_\alpha]) \propto \exp\left\{\varepsilon^{-1}\lambda_\alpha(x[S_\alpha]) + \sum_{j:C_j\subseteq S_\alpha}\left(\varepsilon^{-1}\eta_j + \log\varphi_j\right)(x[C_j])\right\}. \tag{20}$$

Comparing with (18) now shows us that[7]

$$\log f_\alpha(x[S_\alpha]) \doteq \left(1+\varepsilon^{-1}\right)\lambda_\alpha(x[S_\alpha]) + \sum_{j:C_j\subseteq S_\alpha}\left(\varepsilon^{-1}\eta_j + \log\varphi_j\right)(x[C_j]). \tag{21}$$

Summing (21) over α now tells us that

$$\left(1+\varepsilon^{-1}\right)\sum_\alpha \lambda_\alpha \doteq \sum_\alpha \log f_\alpha - \sum_j \left(\varepsilon^{-1}\eta_j + \log\varphi_j\right), \tag{22}$$

or equivalently that

$$\varepsilon^{-1}\sum_\alpha \lambda_\alpha \doteq (1+\varepsilon)^{-1}\sum_\alpha \log f_\alpha - (1+\varepsilon)^{-1}\sum_j \left(\varepsilon^{-1}\eta_j + \log\varphi_j\right). \tag{23}$$

Substituting this into (19) gives the stated form (15). □

Remarks 1. A first glance at the form of the solution (15) for g might lead one to believe that by integrating out $x[\sim C_j]$ and using the constraint (14) we will obtain explicitly what η_j is, and therefore have a very concrete representation of the solution. But unfortunately when we integrate over $x[\sim C_j]$ the integration involves the values of η_i for $i \neq j$, and so what we obtain is an *implicit* characterization of the η_i.
2. If we take the expression (15) for the optimal g and formally let ε tend to zero, we find the much simpler expression

$$g(x) \propto \exp\left[\sum_\alpha \log f_\alpha(x[S_\alpha]) + \sum_j \eta_j(x[C_j])\right]. \tag{24}$$

For the reasons explained above, this formal passage to the limit may not deliver an optimal g, and the η_j in any case depend on ε so there are issues there also. Nevertheless, the simple structural form (24) is appealing, and may provide a good place to start looking.

Multivariate Gaussian distributions

A very important special case which can be reduced to a matrix equation is the case where all the distributions are multivariate Gaussian. This is important because of the following result.

[7] The symbol $\doteq$ denotes that the two sides differ by a constant.

Theorem 2 *With the notation of Theorem 1, if the densities f_α and φ_j are all Gaussian, then the multipliers λ_α and η_j are quadratic functions.*

Proof For brevity, write $\eta_j + \varepsilon \log \varphi_j \equiv h_j$. Then rearranging (21) gives us

$$\lambda_\alpha(x[S_\alpha]) \doteq \frac{\varepsilon}{1+\varepsilon} \log f_\alpha(x[S_\alpha]) - \frac{1}{1+\varepsilon} \sum_{\{j:C_j \subseteq S_\alpha\}} h_j(x[C_j]). \tag{25}$$

Rearranging (19) gives us that

$$\begin{aligned} \varepsilon \log g(x) &\doteq \sum_\alpha \lambda_\alpha(x[S_\alpha]) + \sum_j h_j(x[C_j]) \\ &\doteq \frac{\varepsilon}{1+\varepsilon} \left[\sum_\alpha \log f_\alpha(x[S_\alpha]) + \sum_j h_j(x[C_j]) \right], \end{aligned} \tag{26}$$

using (25). Moreover, from (18) using (25) we deduce that

$$\log g_\alpha(x[S_\alpha]) \doteq \frac{1}{1+\varepsilon} \left[\log f_\alpha(x[S_\alpha]) + \sum_{\{j:C_j \subseteq S_\alpha\}} h_j(x[C_j]) \right]. \tag{27}$$

Now g_α is the S_α-marginal of g, so we have

$$\begin{aligned} g_\alpha(x[S_\alpha]) &\propto \exp\left(\frac{1}{1+\varepsilon} \left[\log f_\alpha(x[S_\alpha]) + \sum_{\{j:C_j \subseteq S_\alpha\}} h_j(x[C_j]) \right] \right) \\ &= \int g(x)\, dx[\sim S_\alpha] \\ &\propto \int \exp\left(\frac{1}{1+\varepsilon} \left[\sum_\beta \log f_\beta(x[S_\beta]) + \sum_j h_j(x[C_j]) \right] \right) dx[\sim S_\alpha], \end{aligned}$$

so cancelling out common factors leads to the conclusion that

$$1 \propto \int \exp\left(\frac{1}{1+\varepsilon} \left[\sum_{\beta \neq \alpha} \log f_\beta(x[S_\beta]) + \sum_{C_j \cap S_\alpha = \emptyset} h_j(x[C_j]) \right] \right) dx[\sim S_\alpha]. \tag{28}$$

To spell it out, the right-hand side of (28) does not change as we vary $x[S_\alpha]$. The only place in the right-hand side of (28) where entries of $x[S_\alpha]$ appear is in the sum $\sum_{\beta \neq \alpha} \log f_\beta(x[S_\beta])$, which is a *quadratic* function. All the terms which are second order in $x[S_\alpha]$ can therefore be put to the other side of (28), leaving on the right-hand side only terms linear in $x[S_\alpha]$ in the exponent. We now interpret the right-hand side of (28) as the Laplace transform of some function; the transform is exponential-quadratic, so the transformed function must be exponential-quadratic. This tells us that the h_j must be quadratic functions, and therefore that the η_j are quadratic, since the $\log \varphi_j$ are quadratic by hypothesis. Immediately from (25) we see that the λ_α are also quadratic functions. □

So we suppose that model M^α states that the law of $Y[S_\alpha]$ will be $N(b_\alpha, \Sigma_\alpha)$, and the tile marginal for C_j is also a multivariate Gaussian $N(a_j, Q_j)$. Since we demand that the tile marginals of g are the laws q^j, there is no freedom for the means of g, so without much loss of generality we shall for clarity of exposition suppose that the b_α and the a_j are all *zero*, leaving us to consider only centred Gaussians. The multipliers η_j will also be centred Gaussians,

$$\eta_j(x[C_j]) = \tfrac{1}{2}\, x[C_j] \cdot W_j\, x[C_j] \tag{29}$$

for some symmetric (though not necessarily positive-definite) matrices W_j. The form (15) of the optimal g will now give us

$$2(1+\varepsilon)\log g(x) \doteq -\sum_\alpha x[S_\alpha]\cdot\Sigma_\alpha^{-1}x[S_\alpha]+\sum_j x[C_j]\cdot\left(W_j-\varepsilon Q_j^{-1}\right)x[C_j]$$
$$=-x\cdot\mathbf{V}x+x\cdot\mathbf{W}x, \tag{30}$$

where we write $\mathbf{W}$ for the unknown block-diagonal matrix of the W_j, and $\mathbf{V}$ for the remaining (positive-definite) terms of the quadratic form. It is now clear that we must choose the W_j so that *for all j the C_j-diagonal block of $(1+\varepsilon)(\mathbf{V}-\mathbf{W})^{-1}$ is the given covariance matrix Q_j*. Determining the W_j given the Σ_α and the Q_j appears to be a challenging numerical problem. The approach used in the following numerical study minimized the relative entropy numerically, but a direct method would be good.

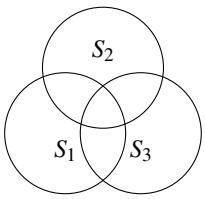
Example 2 Revisited Example 2 fits into the multivariate Gaussian framework just presented; let us see how the general results look in this case. We have that each Q_j is the 1×1 matrix 1, and each Σ_α is the 2×2 identity matrix. Hence

$$V=\begin{pmatrix}1+\varepsilon & 0 & 0\\ 0 & 2+\varepsilon & 0\\ 0 & 0 & 1+\varepsilon\end{pmatrix},\qquad W=\begin{pmatrix}W_1 & 0 & 0\\ 0 & W_2 & 0\\ 0 & 0 & W_3\end{pmatrix}.$$

In order to match the C_j-marginals for each tile, we see that we must take $W_1=W_3=0$, and $W_2=1$, values which do not depend on ε. The solution picked out from the family (9) is the one where X_1 and X_3 are independent, perhaps not surprisingly.

A numerical study

This study worked with 10 years of daily price data on 35 major US stocks, and proposed three models, each making statements about exactly 20 of the stocks. In each of the non-empty subsets of the Venn diagram below there were 5 stocks, the allocation of stocks to subsets being arbitrary. Without going into the exact details, each model generated a predictive distribution each day, and the Bayesian combination was used to come up with a predictive distribution for each equivalence class C_j of variables.



This predictive distribution was of course a mixture distribution, in fact, a mixture of multivariate Gaussians. The combination of the predictive distributions was done by pretending that each predictive distribution was a multivariate Gaussian with matching mean and covariance, and then using the results of the previous subsection to come up with an overall mean $\bar{\mu}$ and covariance $\bar{V}$ for the predicted distribution of all 35 asset returns. Then the portfolio of the assets used was simply $\bar{V}^{-1}\bar{\mu}$. Not very much can be concluded from the output Fig. 1, except that the realized gains seem to have behaved quite sensibly, without any large jumps or swings. The allocation of assets to models was quite arbitrary, so we would not expect that anything particularly good would result; with some more guidance over the choice of

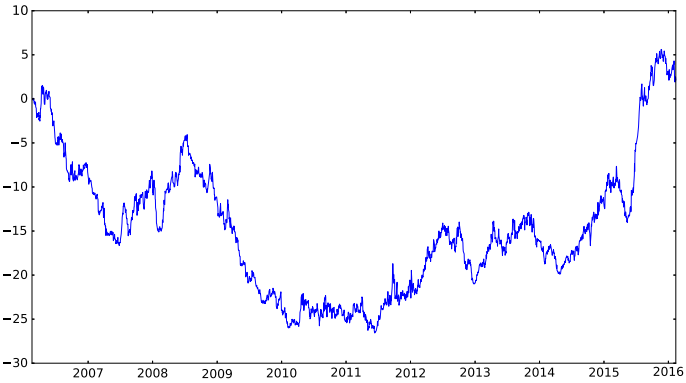


Fig. 1 Cumulative P&L of mean-variance strategy using the combination predictive distribution

variables for models, we may well be able to produce some more interesting P&L plots, but the point to note is that this methodology does give a way to combine small models into a sensible algorithm for dealing with a much larger set of assets, which was the goal of the study.

4 Conclusions

The problem of making a good portfolio from a large number of assets is an important and challenging one; this paper offers an approach that envisages that the big problem is first broken down into smaller more manageable problems. The key feature is that we do not need to suppose that the smaller problems make statements about *disjoint* sets of assets, but rather that understanding of co-dependence of assets can come from multiple separate models which each embody some part of the co-dependence of different assets.

The given models naturally partition the assets into equivalence classes (tiles) on which standard Bayesian model averaging can be applied. We have developed an entropy-minimization method of combining the measures on different assets into a consensus measure. Performing the optimization under constraints on the marginal laws on the individual tiles leads us to the overall combination.

References

1. Geweke, J., Amisano, G.: Optimal prediction pools. J. Econom. **164**(1), 130–141 (2011)
2. Fan, J., Liao, Y., Liu, H.: An overview of the estimation of large covariance and precision matrices. Econom. J. **19**(1), C1–C32 (2016)
3. Bates, J.M., Granger, C.W.J.: Combination of forecasts. Oper. Res. Q. **20**(4), 451–468 (1969)
4. Elliott, G., Timmermann, A.: Economic Forecasting. Princeton University Press, Princeton (2016)

5. Pettenuzzo, D., Ravazzolo, F.: Optimal portfolio choice under decision-based model combinations. J. Appl. Econom. **31**(7), 1312–1332 (2016)
6. Schachermayer, W., Beiglböck, M., Veliyev, B.: A short proof of the Doob–Meyer theorem. Stoch. Process. Appl. **122**(4), 1204–1209 (2012)