“Reverse engineering” research portfolio synergies and tradeoffs from domain expertise in minimum data contexts

Benjamin Schiek [1] *

1 Alliance Bioversity-CIAT, Km 17 Recta Cali-Palmira, Valle del Cauca, Colombia, CP 763537

* Corresponding author: b.schiek@cgiar.org

Abstract

In research portfolio planning contexts, an estimate of research policy and project synergies/tradeoffs (i.e. covariances) is essential to the optimal leveraging of institution resources. The data by which to make such estimates generally do not exist. Research institutions may often draw on domain expertise to fill this gap, but it is not clear how such ad hoc information can be quantified and fed into an optimal resource allocation workflow. Drawing on principal components analysis, I propose a method for “reverse engineering” synergies/tradeoffs from domain expertise at both the policy and project level. I discuss extentions to other problems and detail how the method can be fed into a research portfolio optimization workflow.

Introduction

Agricultural research for development (AR4D) institutions tend to give careful consideration to the formulation of their policies and strategic objectives, but very little, if any, consideration to the tradeoffs and synergies that may arise between policies. An institution may decide to simultaneously pursue, for example, food security and environmental sustainability as overarching strategic objectives, without considering the implicit tradeoffs between capital-intensive, high input agriculture, on the one hand, and pro-poor, agroecological agriculture, on the other. Such tradeoffs mean that progress towards one strategic objective (SO) can offset or even annul progress towards another. Conversely, there may be areas where the institution's policies complement each other, generating synergies and enhancing impacts.

A parallel problem exists at the project level: careful consideration is often given to the potential impacts of individual research projects within the institution's portfolio; but very little, if any, consideration is given to the tradeoffs and synergies that may arise between projects. AR4D institutions can usually draw on a wealth of domain expertise to shed light on these synergies and tradeoffs in a piecemeal fashion; but efforts to scale and quantify such ad hoc assessments—for example, through the Delphi Method or the Analytical Hierarchy Process [1]—are costly and time consuming. There are also inevitably gaps where domain experts are unable or hesitant to venture an estimate. For example: What is the synergy/tradeoff between a heat tolerant bean project and a digital agriculture project?

In this article, I propose a low cost, expedient method for "reverse engineering" synergies and tradeoffs at both the policy and project levels. Drawing on principal components analysis, I show how a project synergies/tradeoffs (a.k.a. covariance) matrix can be approximated based upon an expert survey of correlations between the institution's projects and its policies (or SOs). It turns out that the project level problem is mathematically dual to the policy level problem, such that a policy synergies/tradeoffs matrix is also obtained in this process.

To build intuition and provide a proof of concept, I illustrate the reverse engineering method with a graphical example based on financial data. I then walk through an illustrative example of how the method applies in the AR4D context. I then discuss potential applications in plant breeding and research portfolio optimization.

"Reverse engineering" principal components analysis to deduce synergies and tradeoffs

Signals from noise: dimensional reduction of portfolios

In principal components analysis, a dataset X containing τ observations of n variables is distilled into a dataset S of just $m < n$ variables that capture the main tendencies and structure in the data.[1] The distilled matrix S is defined

$$S = X\tilde{P} \tag{1}$$

Where $\tilde{P}$ is a matrix containing the m leading eigenvectors of the full set of eigenvectors P, which is taken from the eigendecomposition of the data covariance matrix Σ_{XX} (equation 2).

[1] The data are always centered. If the variables in X follow diverse scalings and/or units of measurement (i.e. if apples are being compared to oranges), then X should also be scaled to unit variance. In this exposition, the variables are all of the same type, and so X is centered but not scaled. See Abdi [2] for an introduction to principal components analysis.

$$\Sigma_{XX} = P\Gamma P' \tag{2}$$

Where Γ is the diagonal matrix of the eigenvalues of Σ_{XX}.

From the definition (1), it follows that the columns of S are uncorrelated with each other, and that their variance is given by the leading m eigenvalues of the data covariance matrix.

$$\begin{aligned}
\Sigma_{SS} &= \frac{1}{n-1} S'S \\
&= \frac{1}{n-1} \tilde{P}' X' X \tilde{P} \\
&= \tilde{P}' \Sigma_{XX} \tilde{P} \\
&= \tilde{P}' P \Gamma P' \tilde{P} = \tilde{\Gamma}
\end{aligned} \tag{3}$$

The columns of the distilled matrix S are often referred to as the principal components (PC), or the PC scores, or the factor scores. When dealing with noisy time series, as in this article, they might just as well be referred to as the "signals", in the sense that they are signals extracted from noise.

There then remains the question of what essential process these dimensions or signals S describe. This can be interpreted based on how correlated they are with the variables in X. These correlations (K_{XS}) are found by first finding their corresponding covariances (Σ_{XS}).

$$\begin{aligned}
\Sigma_{XS} &= \frac{1}{n-1} X'S \\
&= \frac{1}{n-1} X' X \tilde{P} \\
&= K\tilde{P} = P\Gamma P' \tilde{P} \\
&= \tilde{P}\tilde{\Gamma}
\end{aligned} \tag{4}$$

Given the standard deviations $\boldsymbol{\sigma}_X$, the correlation matrix K_{XS} then follows as

$$\begin{aligned}
K_{XS} &= D(\boldsymbol{\sigma}_X)^{-1} \Sigma_{XS} D(\boldsymbol{\sigma}_S)^{-1} \\
&= D(\boldsymbol{\sigma}_X)^{-1} \tilde{P}\tilde{\Gamma} D(\boldsymbol{\sigma}_S)^{-1}
\end{aligned} \tag{5}$$

(Where the notation $D(\boldsymbol{\sigma}_X)$ stands for a diagonal matrix with the vector $\boldsymbol{\sigma}_X$ along the diagonal.) But the standard deviations of the signals are just the square roots of the retained eigenvalues (recall equation 3), so this reduces to

$$\begin{aligned}
K_{XS} &= D(\boldsymbol{\sigma}_X)^{-1} \Sigma_{XS} \tilde{\Gamma}^{-\frac{1}{2}} \\
&= D(\boldsymbol{\sigma}_X)^{-1} \tilde{P}\tilde{\Gamma}\tilde{\Gamma}^{-\frac{1}{2}} \\
&= D(\boldsymbol{\sigma}_X)^{-1} \tilde{P}\tilde{\Gamma}^{\frac{1}{2}}
\end{aligned} \tag{6}$$

Note that if X were scaled to unit variance, then this would reduce further to

$$K_{XS} = \tilde{P}\tilde{\Gamma}^{\frac{1}{2}} \tag{7}$$

The correlations matrix K_{XS} is sometimes referred to as the "loadings" matrix, in the sense that it indicates how much each variable in X loads onto a given signal (or, vice versa, how much each signal loads onto a given variable).[2] In keeping with this convention, and in order to reduce notational clutter, K_{XS} is henceforth relabeled L.

An example of loadings is given in Fig **??**. In this case, the variables are the daily returns of 11 financial securities covering the period 2019-01-29 to 2019-04-30.[3] The signals are presented in descending order of their corresponding eigenvalues, with Signal 1 representing the principal component with the highest eigenvalue. The eigenvalue reflects the degree to which the signal describes the overall evolution of the data. Here, only the first four signals of the financial data set are shown. The question of how many signals should be extracted from the noise is addressed at the end of the section.

Concrete meaning can now be attributed to the otherwise abstract signals by examining the loadings—i.e. by examining how correlated the signals are with the individual price series. Signal 4, for example, appears to have something to do with price movements in Communications, and is negatively correlated with movements in the Real Estate sector. Signal 3, meanwhile, is positively correlated with Real Estate and Utilities, as well as Communications. Signal 3 might thus be loosely characterized as the "Housing and Urban Development" or "HUD" Signal, while Signal 4 might be called, rather convolutedly, the "Telecommunications Not Related to HUD" Signal. The interpretation of Signals 1 and 2 is still less straightforward, since they are both correlated with many portfolio items.

Applying a rotation to clarify loadings

When the loadings are convoluted like this, it is useful to apply an orthogonal rotation to L in order to clarify the picture. That is to say, instead of examining L, one examines $L_{\circlearrowleft}$.

$$L_{\circlearrowleft} = LB \tag{8}$$

Where B is the orthogonal rotation matrix, such that $B'B = I$ and $BB' = I$.

In Fig **??** a special kind of orthogonal rotation, called a varimax rotation, is applied to L. Varimax rotations flesh out structure by maximizing sparseness in the rotated matrix. After applying this rotation, Signal 1 is now clearly representative of Biotechnology and Healthcare, and so might be called the "Pharmaceutical" Signal. Signal 2 loadings are also now more distinctly pronounced, especially Financials, Industrial, and Transportation. Signal 2 might thus be called the "Financial and Physical Infrastructure" Signal. The rotation has also cleared up the overlap between Signals 3 and 4. Signal 4 is now more exclusively descriptive of price movements in the Communications sector and can thus be relabeled, more succinctly, the "Communications" Signal. Likewise, Signal 3 is now more exlusively descriptive of movements in Real Estate and Utilities, with some description of movements in the Consumer Goods sector.[4]

Further visual confirmation of these interpretations of signal meaning is given by plotting the signals in the time domain together with their highest loading portfolio items superimposed (Fig **??**). Note how the highest loading items tend to hew closely to their respective signals.

[2]These terms vary in the literature. Many prefer to call P the loadings.

[3]Downloaded from yahoo finance using the R tidyquant package. The securities chosen for this example are exchange traded funds broadly representative of the U.S. economy. See Appendix for details.

[4]Gopikrishnan, Rosenow, Plerou, and Stanley [3] pursued a similar line of inquiry when they looked at the components of the eigenvectors of a financial data correlation matrix. However, they did not explain that their findings are indicative of PC-asset correlations; nor did they apply an orthogonal rotation to clarify the interpretation.

How many signals to retain?

In practice, the number of signals that should be distilled from the original data set X depends upon how much of the variance in X the researcher wishes to capture or reflect in the signals, and how many signals are required to reach this subjectively determined threshold. The portion of the system's evolution reflected in any given signal (c_i) is defined as the signal's variance divided by the sum of all signal variances. Recalling from equation 3 that a signal's variance is just the corresponding eigenvalue γ_i extracted from Σ_{XX}, this is expressed

$$c_i = \frac{u_i}{\sum_{i=1}^{n} \gamma_i} \tag{9}$$

The cumulative variance captured by a group of k signals is then

$$\kappa = \sum_{i=1}^{k} c_i \tag{10}$$

The individual and cumulative portions explained by each signal are plotted in Fig **??**. Customarily, researchers like to retain signals such that at least 90% of the variance in the original data is explained. The horizontal dashed line in the plot marks this subjective threshold.

The plot shows that, for the financial data set, the leading 6 signals are sufficient to meet this criterion.

Approximating the data correlation matrix from the loadings

An approximate data correlation matrix ($\tilde{K}_{XX}$) can be computed from these retained signals, as the outer product of $L_{\circlearrowleft}$ with itself.

$$\begin{aligned} L_{\circlearrowleft} L'_{\circlearrowleft} &= LB(LB)' = LBB'L' = LL' \\ &= (\tilde{P}\tilde{\Gamma}^{\frac{1}{2}})(\tilde{P}\tilde{\Gamma}^{\frac{1}{2}})' \\ &= \tilde{P}\tilde{\Gamma}^{\frac{1}{2}})\tilde{\Gamma}^{\frac{1}{2}}\tilde{P}' \\ &= \tilde{K}_{XX} \end{aligned} \tag{11}$$

The difference between the financial data correlation matrix and the loadings derived correlation matrix calculated in equation 11 is shown in Fig **??**. Note that the difference is remarkably small for most entries. The signals derived correlation matrix is approximate in the sense that it approximates the data correlation matrix; but it should not necessarily be considered inferior in terms of accuracy. To the extent that the original data are contaminated by noise, the signals derived correlation matrix may prove more accurate with respect to the "true process" that generates the data.

Recovering the signals correlation matrix and leading eigenvectors from the loadings

Note that the orthogonally rotated signals correlation matrix (call this $K_{SS}^{\circlearrowleft}$) can likewise be obtained from the rotated loadings via the inner product $L'_{\circlearrowleft} L_{\circlearrowleft}$.

$$
\begin{aligned}
L_{\circlearrowleft}{}' L_{\circlearrowleft} &= (\tilde{P}\tilde{\Gamma}^{\frac{1}{2}}B)'\tilde{P}\tilde{\Gamma}^{\frac{1}{2}}B \\
&= B'\tilde{\Gamma}^{\frac{1}{2}}\tilde{P}'\tilde{P}\tilde{\Gamma}^{\frac{1}{2}}B \\
&= B'\tilde{\Gamma}B \\
&= K_{SS}^{\circlearrowleft}
\end{aligned}
\tag{12}
$$

The project synergy/tradeoff problem is thus dual to the SO synergy/tradeoff problem.

Note that the unrotated signal variances $\tilde{\Gamma}$ and the rotation matrix B can be recovered via an eigendecomposition of $K_{SS}^{\circlearrowleft}$. Moreover, with $\tilde{\Gamma}$ and B in hand, it is then possible to derive the implicit leading eigenvectors of the data correlation matrix ($\tilde{P}$), as follows.

$$
\tilde{P} = L_{\circlearrowleft}B'\tilde{\Gamma}^{-1/2} \tag{13}
$$

"Reverse engineering" project and SO correlation matrices from domain knowledge

The foregoing implies that it is possible to work backwards from the orthogonally rotated loadings $L_{\circlearrowleft}$ to arrive at an approximate data correlation matrix.

In the AR4D context, an institution's SOs are comparable to a set of principal components describing 90% of the problem space that is of interest to the institution. Given a portfolio of projects, a survey of domain experts and/or stakeholders could be conducted to determine how correlated each project is with each SO. These correlations may then be interpreted as orthogonally (not necessarily varimax) rotated loadings $L_{\circlearrowleft}$ corresponding to an unobserved dataset measuring project impact.[5] The project impact correlation matrix can then be calculated from this information via equation 11, and the SO correlation matrix can be calculated via equation 12.

If project risk (standard deviation) can be calculated beforehand during ex-ante impact assessment exercises, then it is straightforward to calculate a project covariance matrix as well. However, risk assessment is still not a standard part of ex-ante impact assessment models.[6] If ex-ante risk assessments are not available, then they can be elicited in the survey of domain experts. Project risk might be crowdsourced, for example, by asking survey participants to estimate the maximum, minimum, and most probable impact of each given project. With these three inputs, it is then straightforward to compute standard deviation on the basis of an assumed project impact probability density.[7]

In this way, project correlation and covariance matrices can be "reverse engineered" from domain knowledge when there is no data. Such an approach makes sense only in contexts where a relative lack of good data is compensated by a relative abundance of good domain knowledge. As a rule of thumb, the appropriateness of this approach may be assessed by meditating upon the conceptual ratio ν.

[5] It is important to have a clear notion of this unobserved dataset. When two projects are said to be correlated, what exactly about the two projects is correlated? In many cases, project return on investment (ROI) may be the variable of interest. But other measures might also be considered, such as net present value or scaleability.

[6] Alston and Norton acknowledged in 1995 that the treatment of risk in impact assessment models was "rudimentary and in need of further refinement" [4]. Unfortunately, this remains true today.

[7] For example, the minimum and maximum could be interpreted as the bounds of the 95% confidence interval of a lognormal probability density, and the "most probable impact" could be interpreted as its mode. From this it is then straightforward to derive the standard deviation.

$$\nu = \frac{\text{confidence in domain knowledge}}{\text{confidence in data}}$$

As ν is higher, the reverse engineering approach makes more sense. As ν is lower, it becomes more appropriate to estimate the covariance matrix on the basis of data. For values of ν close to 1, a mixture of the two approaches might be considered. By this measure, financial contexts are an inappropriate setting for the method proposed here, whereas AR4D contexts are appropriate.

An illustrative example

In the example below, a hypothetical AR4D institution has the task of identifying synergies and tradeoffs in its project portfolio; and is also interested in quantifying any synergies and tradeoffs between its overarching policies. The institution's projects are listed in Table 1. The projects are loosely grouped into four categories to facilitate interpretation of the subsequent graphics, but there is no strict rule followed, and clearly some overlap, in the grouping.

Table 1. Hypothetical list of AR4D projects

Project	Group
Mega Maize	High Value Yield Enhancement
Hyper Rice	High Value Yield Enhancement
Ultra Cow	High Value Yield Enhancement
Cassava for Bio-ethanol	Smallholder Resilience
Triple Purpose Sweet Potato	Smallholder Resilience
Dairy Cooperative	Smallholder Resilience
Multi-stakeholder Platforms	Smallholder Resilience
Heat Tolerant Beans	Climate Smart Agriculture
Coffee Agroforestry	Climate Smart Agriculture
Digital Agriculture	Climate Smart Agriculture
Low Emission Silvopastoral	Climate Smart Agriculture

The institution's policies or SOs in this example are "Economic Growth", "Income Equality", "Environmental Sustainability", and "Nutritional Security", which roughly correspond to UN Sustainable Development Goals 8, 1, 13, and 3, respectively. Project-SO correlations are elicited via a survey of domain experts and/or stakeholders. Literature may also be consulted.

It should be clearly explained to survey participants that a positive project-signal correlation means the project contributes toward the SO (i.e. is a synergy), while a negative correlation means the project works against it (i.e. is a tradeoff); and a correlation of zero means that the project has no influence upon the given SO one way or the other. The language used in this survey should be familiar to participants. In most AR4D resource allocation settings, what I characterized in the financial example above as "signals" should probably be referred to as "policies", "strategic objectives", "criteria", or simply "goals".

Survey participants should also be encouraged to keep in mind that no AR4D project can "be all things to all people". A new yield enhancing variety of a high value crop, for example, might contribute towards increased trade competitiveness and GDP growth, but at the cost of increased deforestation and use of chemical inputs that degrade the environment. Conversely, a climate smart or pro-poor AR4D proposal might increase long term environmental and socio-economic sustainability at the cost of reduced short-medium term growth and competitiveness. These tradeoffs require careful consideration.[8] The results of the survey are summarized in Fig **??**.

The survey exercise concludes. The resulting crowdsourced project-SO correlations are then interpreted as the orthogonally rotated loadings matrix $L_{\circlearrowleft}$. The project correlation matrix is then reverse engineered from $L_{\circlearrowleft}$ via equation 11, while the SO correlation matrix is reverse engineered via equation 12.[9]

The project correlation matrix (Fig **??**) can then be used to orient stakeholder discussions regarding tradeoffs and synergies between projects. Some of the matrix elements may serve to confirm expectations, while other elements may come as a surprise, or serve to fill in a gap where experts are hesitant to venture an estimate. It probably comes as no surprise to the hypothetical survey participants, for example, that the high yielding, high value AR4D projects (Hyper Rice, Mega Maize, and Ultra Cow) are strongly correlated with each other, or that they are negatively correlated with some of the climate smart projects (the Low Emission Silvopastoral proposal, in particular). On the other hand, few experts would be willing to venture an assessment of the synergy or tradeoff between the Cassava for Bio-ethanol and Coffee Agroforestry projects. The deduced covariance matrix effectively fills in such gaps with values that maximize consistency with the domain knowledge captured by the survey.

Likewise, the SO correlation matrix (Fig **??**) can be useful in orienting discussion regarding tradeoffs and synergies between SOs. For example, the matrix indicates that enhanced impacts resulting from synergies between the Economic Equality SO and the Environmental Sustainability and Nutritional Security SOs is partially offset by a tradeoff between the Environmental Sustainability and Nutritional Security SOs. Moreover, it may come as a surprise that very little tradeoff exists between the GDP Growth and Economic Equality SOs. The reverse engineered matrix thus arms the institution with a rough guide by which to capitalize on synergies while mitigating tradeoffs.

Discussion

A covariance matrix reverse engineered from domain knowledge in the manner proposed above offers a perspective on otherwise unquantifiable project and SO synergies and tradeoffs. The accuracy of this perspective depends on 1) how completely the chosen SOs capture the evolution of projects within the problem space (in the literal sense of equation 10); and 2) the accuracy of the domain knowledge whence SO-project correlations are deduced. It is thus imporant to apply this method in contexts where there is a high degree of confidence in domain knowledge compensating a general lack of good data (i.e. a high ν ratio). The method is open to criticism insofar as the domain knowledge is skewed by institutional inertia, politicized thinking, and other sources of subjective bias. However, the alternative method of covariance estimation based on data does not necessarily have a comparative advantage in this respect, as it is likewise subject to a host of different, but no less problematic, sources of bias and error.

Regardless of accuracy, the proposed method may have value as a consensus building tool regarding synergies and tradeoffs about which expert opinions differ or are lacking altogether. The method fills in such gaps with the values that effectively maximize

[8] Participants might also be encouraged to beware of any received wisdom regarding tradeoffs and synergies. For example, it is customary in AR4D communities to assume that economic growth and economic equality are mutually exclusive goals [4], whereas recent empirical research suggests a much more nuanced and synergistic relation [5].

[9] Since these are correlation matrices, their diagonal elements must equal 1. When deduced from a set of $m < n$ loadings, the diagonal elements will diverge from 1 (whether crowd- or data-sourced). In the correlation matrices below, then, I correct for this divergence by dividing the matrices through by their diagonals.

consistency with the expert knowledge captured by the survey. In this process, the method may confront experts and stakeholders with potentially surprising logical implications of what they (think they) know about the problem space, and about the evolution of projects and policies through that space, thereby stimulating policy debate and dialogue.

Potential application in plant breeding decision pipelines

Applications of the method presented here are not limited to the assessment of policies and projects. Another potential area of application within the AR4D arena, for example, is in the assessment of trait and variety correlations.

Plant breeders are typically tasked with the development of new varieties featuring a particular new trait—say, for example, resistance to a particular pest or disease—as well as numerous other traits such as fast maturation time, a particular taste, color, shape, nutritional content, and so on. In this process, a map of synergies and tradeoffs between traits and between varieties may be useful in guiding selection decisions.

In this setting, varieties play the role that projects do in the previous example, while traits are analogous to the set of principal components describing 90% of the problem space. Correlations between varieties and traits are elicited through a survey of breeding experts. A hypothetical example of such a crowdsourcing exercise for beans is given in Fig **??**. The variety and trait correlation matrices are then reverse engineered from this information in Fig **??**.

Potential application in research portfolio optimization

Some may be tempted to use the reverse engineered covariance matrix in a risk-adjusted portfolio optimization problem, so as to solve for the optimal resource allocation across projects. For example, if expected (logged) portfolio utility is defined

$$E[\ln(U)] = E[\mathbf{x}]' \ln(\mathbf{w}) \tag{14}$$

where $\mathbf{x}$ is the vector of project returns (i.e. the percentage increase in portfolio utility per one percent increase in project funding) and $\mathbf{w}$ is the vector of budget allocations invested in each project, then portfolio variance or risk, it follows, is defined

$$Var[\ln(U)] = \ln(\mathbf{w})' \Sigma_{XX} \ln(\mathbf{w}) \tag{15}$$

In the absence of data by which to calculate the project covariances in Σ_{XX}, an AR4D institution may try to substitute the reverse engineered covariance matrix $\tilde{\Sigma}_{XX}$, and then solve the problem

$$\min_{\mathbf{w}} \ln(\mathbf{w})' \tilde{\Sigma}_{XX} \ln(\mathbf{w}) \quad s.t. \quad E[\mathbf{x}]' \ln(\mathbf{w}) = E[U_R] \quad \mathbf{1}' \ln(\mathbf{w}) = U_C \tag{16}$$

where U_R and U_C are the institution's return target and budget constraint, respectively, and $\mathbf{1}$ is a vector of ones. However, for this problem to be well posed, the covariance matrix must be invertible. The reverse engineered covariance matrix has $n - m$ eigenvalues equal to zero, and so is not invertible. The constrained risk minimization problem in equation 16 is thus ill posed.

On the other hand, the reverse engineered policy corrrelation matrix $K_{SS}^{\circlearrowleft}$ (equation 12) *is* invertible, thereby opening up the possibility of solving for optimal "policy weights"—i.e. the weight or emphasis given to each policy by the institution. Such weights are often assigned in a highly subjective, ad hoc manner. The method pursued thus far suggests the following, more rigorous approach.

In this case, $\mathbf{x}$ stands for the "policy returns", i.e. the institution's returns under each SO, and the $\mathbf{w}$ stand for the amount invested under each policy. Moreover, $Var(\ln(U))$ is replaced by the quantity $\ln(\mathbf{w})' K_{SS}^{\circlearrowleft} \ln(\mathbf{w})$. This is a nuanced quantity. Because the policy variances are scaled to unity in the correlation matrix $K_{SS}^{\circlearrowleft}$, it is less a reflection of portfolio risk than it is an indicator of portfolio net synergy, i.e., total synergy minus total tradeoffs, given an investment allocation $\mathbf{w}$. Since net synergy is something desireable, the problem becomes a constrained synergy maximization problem, as opposed to a constrained risk minimization problem.

$$\max_{\mathbf{w}} \ln(\mathbf{w})' K_{SS}^{\circlearrowleft} \ln(\mathbf{w}) \quad s.t. \quad E[\mathbf{x}]' \ln(\mathbf{w}) = U_R \quad \mathbf{1}' \ln(\mathbf{w}) = U_C \tag{17}$$

The Lagrangian and first order conditions are then:

$$\begin{aligned} \mathcal{L} &= \ln(\mathbf{w})' K_{SS}^{\circlearrowleft} \ln(\mathbf{w}) + \lambda_R (E[\mathbf{x}]' \ln(\mathbf{w}) - U_R) - \lambda_C (\mathbf{1}' \ln(\mathbf{w}) - U_C) \\ \nabla_{\ln(\mathbf{x})} \mathcal{L} &= 2K_{SS} \ln(\mathbf{w}) + \lambda_R E[\mathbf{x}] - \lambda_C \mathbf{1} = \mathbf{0} \end{aligned} \tag{18}$$

Where $\mathbf{0}$ is a vector of zeroes. The second order condition then follows as

$$\ln(\mathbf{w})' \nabla^2_{\ln(\mathbf{w})} \mathcal{L} \ln(\mathbf{w}) = \ln(\mathbf{w})' K_{SS}^{\circlearrowleft} \ln(\mathbf{w}) < 0 \tag{19}$$

The first order conditions can then be solved for the optimal policy weights as follows.

$$\mathbf{w}^* = 1/2 K_{SS}^{\circlearrowleft -1} [\mathbf{x}, \mathbf{1}] \begin{bmatrix} -\lambda_R \\ \lambda_B \end{bmatrix} \tag{20}$$

where

$$\begin{bmatrix} -\lambda_R \\ \lambda_B \end{bmatrix} = 2M^{-1} \begin{bmatrix} U_R \\ U_B \end{bmatrix} \; ; \quad M = [\mathbf{x}, \mathbf{1}]' K_{SS}^{\circlearrowleft -1} [\mathbf{x}, \mathbf{1}] \tag{21}$$

Solving for $\mathbf{w}^*$ requires not only the reverse engineered correlation matrix, but also an estimate of the expected policy returns $E[\mathbf{x}]$. In high ν contexts, the data required to arrive at such an estimate do not readily exist (for the same reasons that the data needed to compute a covariance matrix do not exist). However, domain expertise in such circumstances is typically sufficient to approximate $E[\mathbf{x}]$ through ex-ante impact assessment studies.

The optimal resource allocation $\mathbf{w}^*$ may refer to actual funds, or may be interpreted more loosely as an allocation of attention, time, political will, or enthusiasm, as appropriate. In the case where $K_{SS}^{\circlearrowleft}$ is a variety correlation matrix, $\mathbf{w}^*$ represents the optimal resource allocation across the traits under consideration.

Conclusion

For a long time now, research institutions have faced increasing donor pressure to "do more with less" [6], "prove their relevance" [1], "show value for money" [7], and otherwise demonstrate "more efficient spending of resources" [8].

In response to this pressure, researchers have focused on the development of models for the ex-ante impact assessment of individual projects [4,9–11]. However, new decision support tools are still urgently required at the portfolio level to determine optimal resource allocations across strategic objectives. In the absence of such tools, resource allocation procedures have been repeatedly undercut by stakeholder politics, institutional inertia, and other forms of subjective bias; and this, in turn, has contributed to an historic level of toxicity in AR4D donor-researcher relations [12–14]. The toxicity is palpable across other disciplines as well [15,16].

The task of allocating limited resources across strategic objectives that are all, in one way or another, vitally important, will never be an easy one. Nonetheless, it stands to reason that the introduction of objective, transparent resource allocation mechanisms can substantially ameliorate the current atmosphere of distrust. As a step in this direction, above I have presented a novel project and policy synergy/tradeoff reverse engineering method based on principal components analysis. The proposed method aids in identifying areas in the AR4D portfolio where research impacts capitalize upon and enhance, or, conversely, annul and offset, each other.

The method can be applied to portfolios of projects or portfolios of policies. For policy portfolios, I showed how the reverse engineered synergy tradeoff matrix may be used in a constrained portfolio synergy maximization problem to solve for optimal policy weights—which can then, in turn, guide the optimal allocation of institution resources. I have also sketched out how the proposed method might be applied to the analogous problem of plant trait selection. The proposed method is not limited to these expository examples, nor even to the AR4D context, but rather applies to any portfolio level planning context where a relative lack of data is compensated by a relative abundance of domain expertise.

References

1. Braunschweig T, Becker B. Choosing research priorities by using the analytic hierarchy process: An application to international agriculture. R&D Management. Wiley Online Library; 2004;34: 77–86.

2. Abdi H, Williams LJ. Principal component analysis. Wiley interdisciplinary reviews: computational statistics. Wiley Online Library; 2010;2: 433–459.

3. Gopikrishnan P, Rosenow B, Plerou V, Stanley HE. Quantifying and interpreting collective behavior in financial markets. Physical Review E. APS; 2001;64: 035106.

4. Alston JM, Norton GW. Science Under Scarcity: Principles and Practice for Agricultural Research Evaluation and Priority Setting [Internet]. Cornell University Press; 1995. pp. 1064–1065. doi:10.1016/S0169-5150(96)01197-8

5. Berg AG, Ostry JD. Inequality and unsustainable growth: Two sides of the same coin? IMF Economic Review. Springer; 2017;65: 792–815.

6. Norton GW, Pardey PG, Alston JM. Economic Issues in Agricultural Research Priority Setting. American Journal of Agricultural Economics. Oxford University PressAgricultural & Applied Economics Association; 1992;74: 1089–1094. doi:10.2307/1242762

7. Yet B, Constantinou A, Fenton N, Neil M, Luedeling E, Shepherd K. A bayesian network framework for project cost, benefit and risk analysis with an agricultural development case study. Expert Systems with Applications. Elsevier; 2016;60: 141–155.

8. Petsakos A, Hareau G, Kleinwechter U, Wiebe K, Sulser TB. Comparing modeling approaches for assessing priorities in international agricultural research. Research Evaluation. Oxford University Press; 2018;27: 145–156.

9. Antle JM. The New Economics of Agriculture. American Journal of Agricultural Economics. 1999;81. doi:10.2307/1244078

10. Antle JM, Valdivia RO, Boote KJ, Janssen S, Jones JW, Porter CH, et al. AgMIP's Transdisciplinary Agricultural Systems Approach to Regional Integrated Assessment of Climate Impacts, Vulnerability, and Adaptation. Handbook of Climate Change and Agroecosystems. 2015; 27–44.

11. Nelson GC, Shively GE. Modeling climate change and agriculture: An introduction to the special issue. Agricultural Economics. Wiley Online Library; 2014;45: 1–2.

12. Birner R, Byerlee D. Synthesis and Lessons Learned from 15 CRP Evaluations [Internet]. Rome, Italy: Independent Evaluation Arrangement (IEA) of CGIAR; 2016. Available: `http://iea.cgiar.org/`

13. Leeuwis C, Klerkx L, Schut M. Reforming the research policy and impact culture in the cgiar: Integrating science and systemic capacity development. Global food security. Elsevier; 2018;16: 17–21.

14. McCalla AF. CGIAR reform-why so difficult? Review, reform, renewal, restructuring, reform again and then" the new cgiar"-so much talk and so little basic structural change-why? 2014;

15. Petsko GA. Herding cats. Science Translational Medicine. 2011;3. doi:DOI: 10.1126/scitranslmed.3002837

16. Moriarty P. Reclaiming academia from post-academia. Nature Nanotechnology. 2008;3: 60–62. doi:10.1038/nnano.2008.11

Appendix: Securities appearing in the financial example

Name	Symbol	Tracks
Financial Select Sector SPDR Fund	XLF	U.S. Financials Sector
Communication Services Select Sector SPDR Fund	XLC	U.S. Communications Sector
Consumer Discretionary Select Sector SPDR Fund	XLY	U.S. Luxury goods Sector
Consumer Staples Select Sector SPDR Fund	XLP	U.S. Consumer goods Sector
Health Care Select Sector SPDR Fund	XLV	U.S. Healthcare Sector
Technology Select Sector SPDR Fund	XLK	U.S. Technology Sector
SPDR Dow Jones REIT ETF	RWR	U.S. Real estate Sector
Utilities Select Sector SPDR Fund	XLU	U.S. Utilities Sector
Industrial Select Sector SPDR Fund	XLI	U.S. Industrial Sector
SPDR S&P Biotech ETF	XBI	U.S. Biotechnology Sector
iShares Transportation Average ETF	IYT	U.S. Transportation Sector