Chapter 15
Mining Matrix Data with Bregman Matrix Divergences for Portfolio Selection

Richard Nock, Brice Magdalou, Eric Briys and Frank Nielsen

15.1 Introduction

If only we always knew ahead of time.... The dream of any stock portfolio manager is to allocate stocks in his portfolio in hindsight so as to always reach maximum wealth. With hindsight, over a given time period, the best strategy is to invest into the best performing stock over that period. However, even this appealing strategy is not without regret. Reallocating everyday to the best stock in hindsight (that is with a perfect sense for ups and downs timing) notwithstanding, Cover has shown that a Constant Rebalancing Portfolio (CRP) strategy can deliver superior results [10]. These superior portfolios have been named Universal Portfolios (UP). In other words, if one follows Cover's advice, a non anticipating portfolio allocation performs (asymptotically) as well as the best constant rebalancing portfolio allocation determined in hindsight. This UP allocation is however not costless as it replicates the payoff, if it existed, of an exotic option, namely a hindsight allocation option. Buying this option, if it were traded, would enable a fund manager to behave as if he always knew everything in hindsight.

Finding useful portfolio allocations, like the CRP allocation, is not however always related to the desire to outperform some pre-agreed benchmark. As Markowitz has shown, investors know that they cannot achieve stock returns greater than the risk-free rate without having to carry some risk [17]. Markowitz designed a decision criterion which, taking both risk and return into account, enables any investor to compute the weights of each individual stock in his preferred portfolio. The investor is assumed to like return but to dislike risk: this is the much celebrated mean-variance

R. Nock (✉) · B. Magdalou · E. Briys
CEREGMIA-Université Antilles-Guyane Martinique, France
e-mail: rnock@martinique.univ-ag.fr

F. Nielsen
Sony CS Labs Inc., Tokyo, Japan
e-mail: nielsen@csl.sony.co.jp

F. Nielsen and R. Bhatia (eds.), *Matrix Information Geometry*,
DOI: 10.1007/978-3-642-30232-9_15,

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

approach to portfolio selection. More specifically, the investor computes the set of efficient portfolios such that the variance of portfolio returns is minimized for a given expected return objective and such that the expected return of the portfolio is maximized for a given variance level. Once, the efficient set is computed, the investor picks his optimal portfolio, namely, that which maximizes his expected utility. This choice process can be simplified if one considers an investor with an exponential utility function and a Gaussian distribution of stock returns. In that case, the optimal portfolio is that which maximizes the spread between the expected return and half the product of variance and the Arrow–Pratt index of absolute risk aversion [23]. Everything goes as if the expected returns were penalized by a quantity that depends both on risk and risk aversion. Although the mean-variance approach has nurtured a rich literature on asset pricing, its main defects are well-known [6, 8]. In particular, it works well in a setting where one can safely assume that returns are governed by a Gaussian distribution. This is a serious limitation that is not supported by empirical data on stock returns.

In the following, we relax this assumption and consider the much broader set of exponential families of distributions. Our first contribution is to show that the mean-variance framework is generalized in this setting by a mean-divergence framework, in which the divergence is a Bregman matrix divergence [7], a class of distortions which generalizes Bregman divergences, that are familiar to machine learning works ([11, 12, 15], and many others). This setting, which is more general than another one studied in the context of finance by the authors with plain Bregman divergences [20], offers a new and general setting (i) to analyze market events and investors' behaviors, as well as a (ii) to design, analyze and test learning algorithms to track efficient portfolios. The divergences we consider are general Bregman matrix divergences that draw upon works in quantum physics [21], as well as a new, even broader class of Bregman matrix divergences whose generator is a combination of functions. This latter class includes as important special case divergences that we call Bregman–Schatten p-divergences, that generalize previous attempts to upgrade p-norms vector divergences to matrices [13]. We analyze risk premia in this general setting. A most interesting finding about the generalization is the fact that the dual affine coordinate systems that stem from the Bregman divergences [2] are those of the *allocations* and *returns* (or wealth). Hence, the general "shape" of the *premium* implicitly establishes a tight bond between these two key components of the (investor, market) pair. Another finding is a *natural market allocation* which pops up in our generalized premium (but simplifies in the mean-variance approach), and defines the optimal but unknown market investment. In the general case, the risk premium thus depends on more than two parameters (the risk aversion parameter and a variance-covariance matrix): it depends on a (convex) premium generator, the investor's allocation, the investor's risk aversion and the natural market allocation. The matrix standpoint on the risk premium reveals the roles of the two main components of allocation matrices: the spectral allocations, *i.e.* the diagonal matrix in the diagonalization of the allocation matrices, and their transition matrices that play as interaction factors between stocks.

Recent papers have directly cast learning in the original mean-variance model, in an on-line learning setting: the objective is to learn and track portfolios exhibiting

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

bounded risk premia over a sequence of market iterations [14, 26]. The setting of these works represents the most direct lineage to our second contribution: the design and analysis, in our mean-divergence model, of an on-line learning algorithm to track *shifting* portfolios of bounded risk premia, which relies upon our Bregman–Schatten p-divergences. Our algorithm is inspired by the popular p-norm algorithms [15]. Given reals $r, \ell > 0$, the algorithm updates symmetric positive definite (SPD) allocations matrices whose r-norm is bounded above by ℓ. The analysis of the algorithm exploits tools from matrix perturbation theory and new properties of Bregman matrix divergences that may be of independent interest. We then provide experiments and comparisons of this algorithm over a period of twelve years of S&P 500 stocks, displaying the ability of the algorithm to track efficient portfolios, and the capacity of the mean-divergence model to spot important events at the market scale, events that would be comparatively dampened in the mean-variance model. Finally, we drill down into a theoretical analysis of our premia, first including a qualitative and quantitative comparison of the matrix divergences we use to others that have been proposed elsewhere [12, 13, 16], and then analyzing the interactions of the two key components of the risk premium: the investor's and the natural market allocations.

The remaining of the paper is organized as follows: Sect. 15.2 presents Bregman matrix divergences and some of their useful properties; Sect. 15.3 presents our generalization of the mean-variance model; Sect. 15.4 analyzes our on-line learning algorithm in our mean-divergence model; Sect. 15.5 presents some experiments; the two last sections respectively discuss further our Bregman matrix divergences with respect to other matrix divergences introduced elsewhere, discuss further the mean-divergence model, and then conclude the paper with avenues for future research.

15.2 Bregman Matrix Divergences

We begin by some definitions. Following [25], capitalized bold letters like $\mathbf{M}$ denote matrices, and italicized bold letters like $\boldsymbol{v}$ denote vectors. Blackboard notations like $\mathbb{S}$ denote subsets of (tuples of, matrices of) reals, and $|\mathbb{S}|$ their cardinal. Calligraphic letters like $\mathcal{A}$ are reserved for algorithms. To make clear notations that rely on economic concepts, we shall use small capitals for them: for example, utility functions are denoted U. The following particular matrices are defined: $\mathbf{I}$, the identity matrix; $\mathbf{Z}$, the all-zero matrix. An allocation matrix $\mathbf{A}$ is SPD; a density matrix is an allocation matrix of unit trace. Unless otherwise explicitly stated in this section and the following ones (Sects. 15.3 and 15.4), matrices are symmetric.

We briefly summarize the extension of Bregman divergences to matrix divergences by using the diagonalization of linear operators [16, 21, 25]. Let ψ be some strictly convex differentiable function whose domain is $\mathrm{dom}(\psi) \subseteq \mathbb{R}$. For any symmetric matrix $\mathbf{N} \in \mathbb{R}^{d\times d}$ whose spectrum satisfies $\mathrm{spec}\,(\mathbf{N}) \subseteq \mathrm{dom}(\psi)$, we let

$$\psi(\mathbf{N}) \doteq \mathrm{Tr}\left(\boldsymbol{\Psi}(\mathbf{N})\right), \quad \boldsymbol{\Psi}(\mathbf{N}) \doteq \sum_{k\geq 0} t_{\psi,k}\mathbf{N}^k, \tag{15.1}$$

Table 15.1 Examples of Bregman matrix divergences. $\boldsymbol{\Sigma}$ is positive definite, $\cdot$ is the Hadamard product, $\boldsymbol{l}, \boldsymbol{n} \in \mathbb{R}^d$ and $\mathbf{1}$ is the all-1 vector

ψ	$D_\psi(\mathbf{L}\|\mathbf{N})$	Comments
$x \log x - x$	$\mathrm{Tr}\,(\mathbf{L}(\log \mathbf{L} - \log \mathbf{N}) - \mathbf{L} + \mathbf{N})$	von Neumann divergence
id.	id. + constraint $\mathrm{Tr}\,(\mathbf{L}) = \mathrm{Tr}\,(\mathbf{N})$	Umegaki's relative entropy [22]
$-\log x$	$\mathrm{Tr}\left(-\log \mathbf{L} + \log \mathbf{N} + \mathbf{L}\mathbf{N}^{-1}\right) - d$	logdet divergence [25]
$x \log x + (1-x)\log(1-x)$	$\mathrm{Tr}\,(\mathbf{L}(\log \mathbf{L} - \log \mathbf{N}) + (\mathbf{I} - \mathbf{L})(\log(\mathbf{I} - \mathbf{L}) - \log(\mathbf{I} - \mathbf{N})))$	binary quantum relative entropy
$x^p \ (p > 1)$	$\mathrm{Tr}\left(\mathbf{L}^p - p\mathbf{L}\mathbf{N}^{p-1} + (p-1)\mathbf{N}^p\right)$	
if $p = 2$	$\mathrm{Tr}\left(\mathbf{L}^2 - 2\mathbf{L}\mathbf{N} + \mathbf{N}^2\right)$	Mahalanobis divergence
	$= (\boldsymbol{l} - \boldsymbol{n})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{l} - \boldsymbol{n})$ if $\mathbf{L} \doteq (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{l})\mathbf{1}^\top \cdot \mathbf{I}$, $\mathbf{N} \doteq (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{n})\mathbf{1}^\top \cdot \mathbf{I}$	
$\log(1 + \exp(x))$	$\mathrm{Tr}\left(\log(\mathbf{I} + \exp(\mathbf{L})) - \log(\mathbf{I} + \exp(\mathbf{N})) - (\mathbf{L} - \mathbf{N})(\mathbf{I} + \exp(\mathbf{N}))^{-1}\exp(\mathbf{N})\right)$	Dual bit entropy
$-\sqrt{1 - x^2}$	$\mathrm{Tr}\left((\mathbf{I} - \mathbf{L}\mathbf{N})(\mathbf{I} - \mathbf{N}^2)^{-1/2} - (\mathbf{I} - \mathbf{L}^2)^{1/2}\right)$	
$\exp(x)$	$\mathrm{Tr}\,(\exp(\mathbf{L}) - (\mathbf{L} - \mathbf{N} + \mathbf{I})\exp(\mathbf{N}))$	
$\phi_p \circ \psi_p \ (p > 1$, Eq. (15.3))	$\frac{1}{2}\,\|\mathbf{L}\|_p^2 - \frac{1}{2}\,\|\mathbf{N}\|_p^2 - \frac{1}{\|\mathbf{N}\|_p^{p-2}}\mathrm{Tr}\left((\mathbf{L} - \mathbf{N})\mathbf{N}\vert\mathbf{N}\vert^{p-2}\right)$	Bregman–Schatten p-divergence

where $t_{\psi,k}$ are the coefficients of a Taylor expansion of ψ, and Tr (.) denotes the trace. A (Bregman) matrix divergence with generator ψ is simply defined as:

$$D_{\psi}(\mathbf{L}\|\mathbf{N}) \doteq \psi(\mathbf{L}) - \psi(\mathbf{N}) - \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\boldsymbol{\nabla}_{\psi}^{\top}(\mathbf{N})\right), \tag{15.2}$$

where $\boldsymbol{\nabla}_{\psi}(\mathbf{N})$ is defined using a Taylor expansion of $\partial\psi/\partial x$, in the same way as $\boldsymbol{\Psi}(\mathbf{N})$ does for ψ in (15.1). We have chosen to provide the definition for the matrix divergence without removing the transpose when $\mathbf{N}$ is symmetric, because it shall be discussed in a general case in Sect. 15.6. Table 15.1 presents some examples of matrix divergences. An interesting and non-trivial extension of matrix divergences, which has not been proposed so far, relies in the functional composition of generators. We define it as follows. For some real-valued functions ϕ and ψ with $\phi \circ \psi$ strictly convex and differentiable, and matrix $\mathbf{N}$, the generator of the divergence is:

$$\phi \circ \psi(\mathbf{N}) \doteq \phi(\psi(\mathbf{N})).$$

Remark that ϕ is computed over the reals. An example of such divergences is of particular relevance: Bregman–Schatten p-divergences, a generalization of the popular Bregman p-norm divergences [15] to symmetric matrices, as follows. Take $\psi_p(x) \doteq |x|^p$, for $p > 1$, and $\phi_p(x) = (1/2)x^{2/p}$. The generator of Bregman–Schatten p-divergence is $\phi_p \circ \psi_p$, and it comes:

$$\phi_p \circ \psi_p(\mathbf{N}) = \frac{1}{2}\|\mathbf{N}\|_p^2 . \tag{15.3}$$

We recall that the Schatten p-norm of a symmetric matrix $\mathbf{N}$ is $\|\mathbf{N}\|_p \doteq \mathrm{Tr}\,(|\mathbf{N}|^p)^{1/p}$, with $|\mathbf{N}| \doteq \mathbf{P}\sqrt{\mathbf{D}^2}\mathbf{P}^\top$, and $\mathbf{P}$ is the (unitary) transition matrix associated to the (diagonal) eigenvalues matrix $\mathbf{D}$. The following Lemma summarizes the main properties of Bregman–Schatten p-divergences, all of which are generalizations of properties known for the usual p-norm divergences. Two reals p and q are said to be Hölder conjugates iff $p, q > 1$ and $(1/p) + (1/q) = 1$.

Lemma 1. *Let p and q be Hölder conjugates, and denote for short*

$$\tilde{\boldsymbol{A}}_p \doteq \boldsymbol{\nabla}_{\phi_p \circ \psi_p}(\boldsymbol{A}). \tag{15.4}$$

The following properties hold true for Bregman–Schatten p-divergences:

$$\tilde{\boldsymbol{N}}_p = \frac{1}{\|\boldsymbol{N}\|_p^{p-2}}\boldsymbol{N}|\boldsymbol{N}|^{p-2}, \tag{15.5}$$

$$\mathrm{Tr}\left(\boldsymbol{N}\tilde{\boldsymbol{N}}_p\right) = \|\boldsymbol{N}\|_p^2 , \tag{15.6}$$

$$\left\|\tilde{\boldsymbol{N}}_q\right\|_p = \|\boldsymbol{N}\|_q , \tag{15.7}$$

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

$$D_{\phi_q \circ \psi_q}(\boldsymbol{L} \| \boldsymbol{N}) = D_{\phi_p \circ \psi_p}(\tilde{\boldsymbol{N}}_q \| \tilde{\boldsymbol{L}}_q). \tag{15.8}$$

Proof sketch: (15.5–15.7) are immediate. To prove (15.8), we prove a relationship of independent interest, namely that $\phi_p \circ \psi_p$ and $\phi_q \circ \psi_q$ are Legendre dual of each other. For any p and q Hölder conjugates, we prove that we have:

$$\widetilde{(\tilde{\mathbf{L}}_q)}_p = \mathbf{L}. \tag{15.9}$$

First, (15.5) brings:

$$\widetilde{(\tilde{\mathbf{L}}_q)}_p = \frac{1}{\left\|\tilde{\mathbf{L}}_q\right\|_p^{p-2}} \tilde{\mathbf{L}}_q |\tilde{\mathbf{L}}_q|^{p-2}. \tag{15.10}$$

We consider separately the terms in (15.10). First, it comes:

$$\begin{aligned}
\left\|\tilde{\mathbf{L}}_q\right\|_p^{p-2} &= \left\|\frac{1}{\|\mathbf{L}\|_q^{q-2}} \mathbf{L}|\mathbf{L}|^{q-2}\right\|_p^{p-2} = \frac{1}{\|\mathbf{L}\|_q^{(p-2)(q-2)}} \mathrm{Tr}\left(|\mathbf{L}|^{(q-1)p}\right)^{\frac{p-2}{p}} \\
&= \frac{1}{\|\mathbf{L}\|_q^{(p-2)(q-2)}} \|\mathbf{L}\|_q^{2-q} = \frac{1}{\|\mathbf{L}\|_q^{(p-1)(q-2)}}.
\end{aligned} \tag{15.11}$$

Then,

$$\begin{aligned}
\tilde{\mathbf{L}}_q |\tilde{\mathbf{L}}_q|^{p-2} &= \frac{1}{\|\mathbf{L}\|_q^{q-2}} \mathbf{L}|\mathbf{L}|^{q-2} \left| \frac{1}{\|\mathbf{L}\|_q^{q-2}} \mathbf{L}|\mathbf{L}|^{q-2} \right|^{p-2} = \frac{1}{\|\mathbf{L}\|_q^{(q-2)(p-1)}} \mathbf{L}|\mathbf{L}|^{qp-q-p} \\
&= \frac{1}{\|\mathbf{L}\|_q^{(q-2)(p-1)}} \mathbf{L},
\end{aligned} \tag{15.12}$$

as indeed $qp - q - p = 0$. Plugging (15.11) and (15.12) into (15.10), one obtains (15.9), as claimed. Then, (15.8) follows from (15.16).

We discuss in Sect. 15.6 a previous definition due to [13] of p-norm matrix divergences, which represents a particular case of Bregman–Schatten p-divergences. The following Lemma, whose proof is omitted to save space, shall be helpful to simplify our proofs, as it avoids the use of rank-4 tensors to bound matrix divergences.

Lemma 2. *Suppose that ϕ is concave, and $\phi \circ \psi$ is strictly convex differentiable. Then $\forall \boldsymbol{L}, \boldsymbol{N}$ two symmetric matrices, there exists $\boldsymbol{U}_\alpha \doteq \alpha \boldsymbol{L} + (1-\alpha)\boldsymbol{N}$ with $\alpha \in [0, 1]$, such that:*

$$D_{\phi \circ \psi}(\boldsymbol{L} \| \boldsymbol{N}) \leq \frac{\nabla_\phi \circ \psi(\boldsymbol{N})}{2} \mathrm{Tr}\left((\boldsymbol{L} - \boldsymbol{N})^2 \left. \frac{\partial^2}{\partial x^2} \psi(x) \right|_{x = \mathbf{U}_\alpha} \right). \tag{15.13}$$

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

Proof We first make a Taylor–Lagrange expansion on ψ; there exists $\alpha \in [0, 1]$ and matrix $\mathbf{U}_\alpha \doteq \alpha\mathbf{L} + (1-\alpha)\mathbf{N}$ for which:

$$\psi(\mathbf{L}) = \psi(\mathbf{N}) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_\psi(\mathbf{N})\right) + \frac{1}{2}\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2 \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right),$$

which implies:

$$\begin{aligned}
&\phi\circ\psi(\mathbf{L})\\
&= \phi\left(\psi(\mathbf{N}) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_\psi(\mathbf{N})\right) + \frac{1}{2}\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2 \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right)\right).
\end{aligned} \tag{15.14}$$

On the other hand, ϕ is concave, and so $\phi(b) \leq \phi(a) + \left.\frac{\partial}{\partial x}\phi(x)\right|_{x=a}(b-a)$. This implies the following upperbound for the right-hand side of (15.14):

$$\begin{aligned}
&\phi\left(\psi(\mathbf{N}) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_\psi(\mathbf{N})\right) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2 \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right)\right)\\
&\leq \phi\circ\psi(\mathbf{N})\\
&\quad + \nabla_\phi\circ\psi(\mathbf{N}) \times \left\{\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_\psi(\mathbf{N})\right) + \frac{1}{2}\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2 \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right)\right\}\\
&= \phi\circ\psi(\mathbf{N}) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_\phi\circ\psi(\mathbf{N})\nabla_\psi(\mathbf{N})\right)\\
&\quad + \frac{1}{2}\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2\nabla_\phi\circ\psi(\mathbf{N}) \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right)\\
&= \phi\circ\psi(\mathbf{N}) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_{\phi\circ\psi}(\mathbf{N})\right)\\
&\quad + \frac{\nabla_\phi\circ\psi(\mathbf{N})}{2}\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2 \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right).
\end{aligned}$$

Putting the resulting inequality into (15.14) yields:

$$\begin{aligned}
\phi\circ\psi(\mathbf{L}) \leq{}& \phi\circ\psi(\mathbf{N}) + \mathrm{Tr}\left((\mathbf{L}-\mathbf{N})\nabla_{\phi\circ\psi}(\mathbf{N})\right)\\
&+ \frac{\nabla_\phi\circ\psi(\mathbf{N})}{2}\mathrm{Tr}\left((\mathbf{L}-\mathbf{N})^2 \left.\frac{\partial^2}{\partial x^2}\psi(x)\right|_{x=\mathbf{U}\alpha}\right).
\end{aligned}$$

Rearranging and introducing Bregman matrix divergences, we obtain (15.13), as claimed. □

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

15.3 Mean-Sivergence: A Generalization of Markowitz' Mean-Variance Model

Our generalization is in fact two-way as it relaxes both the normal assumption and the vector-based allocations of the original model. It is encapsulated by regular exponential families [4] with matrix supports, as follows. We first define the matrix Legendre dual of strictly convex differentiable ψ as:

$$\psi^\star(\tilde{\mathbf{N}}) \doteq \sup_{\mathrm{spec}(\mathbf{N})\subset\mathrm{dom}(\psi)} \{\mathrm{Tr}\left(\mathbf{N}\tilde{\mathbf{N}}^\top\right) - \psi(\mathbf{N})\}. \tag{15.15}$$

We can easily find the exact expression for $\psi^\star$. Indeed, $\tilde{\mathbf{N}} = \nabla_\psi(\mathbf{N})$, and thus $\psi^\star(\tilde{\mathbf{N}}) = \mathrm{Tr}\left(\nabla_\psi^{-1}(\tilde{\mathbf{N}})\tilde{\mathbf{N}}^\top\right) - \psi(\nabla_\psi^{-1}(\tilde{\mathbf{N}}))$, out of which it comes:

$$D_\psi(\mathbf{L}\|\mathbf{N}) = \psi(\mathbf{L}) + \psi^\star(\tilde{\mathbf{N}}) - \mathrm{Tr}\left(\mathbf{L}\tilde{\mathbf{N}}^\top\right) = D_{\psi^\star}(\nabla_\psi(\mathbf{N})\|\nabla_\psi(\mathbf{L})). \tag{15.16}$$

Let $\mathbf{W}$ model a stochastic behavior of the market such that, given $\mathbf{A}$ an allocation matrix, the quantity

$$\omega^F \doteq \mathrm{Tr}\left(\mathbf{A}\mathbf{W}^\top\right) \tag{15.17}$$

models the *wealth* (or reward) retrieved from the Market. In what follows, $\mathbf{W}$ models market returns, and satisfies spec $(\mathbf{W}) \subset [-1, +\infty)$. The stochastic behavior of the market comes from the choice of $\mathbf{W}$ according to regular exponential families [4] using matrix divergences, as follows:

$$\begin{aligned} p_\psi(\mathbf{W}; \boldsymbol{\Theta}) &\doteq \exp\left(\mathrm{Tr}\left(\boldsymbol{\Theta}\mathbf{W}^\top\right) - \psi(\boldsymbol{\Theta})\right) b(\mathbf{W}) && (15.18)\\ &= \exp\left(-D_{\psi^\star}(\mathbf{W}\|\nabla_\psi(\boldsymbol{\Theta})) + \psi^\star(\mathbf{W})\right) b(\mathbf{W}), && (15.19)\end{aligned}$$

where $\boldsymbol{\Theta}$ defines the natural matrix parameter of the family and (15.19) follows from (15.16) [4]. Up to a normalization factor which does not depend on $\boldsymbol{\Theta}$, this density is in fact proportional to a ratio of two determinants:

$$p_\psi(\mathbf{W}; \boldsymbol{\Theta}) \propto \frac{\det\exp(\mathbf{W}\boldsymbol{\Theta}^\top)}{\det\exp(\boldsymbol{\Psi}(\boldsymbol{\Theta}))}. \tag{15.20}$$

It is not hard to see that the following holds true for p_ψ defined as in (15.19):

$$\nabla_\psi(\boldsymbol{\Theta}) = \mathrm{E}_{\mathbf{W}\sim p_\psi}[\mathbf{W}], \tag{15.21}$$

with $\mathrm{E}_{.}[.]$ the expectation. Equation (15.21) establishes the connection between natural parameters and expectation parameters for the exponential families we consider [2]. It also allows to make a useful parallel between $\mathrm{Tr}\left(\boldsymbol{\Theta}\mathbf{W}^\top\right)$ in the general setting (15.18) and ω^F in our application (15.17): while the expectation parameters model the average market returns, the natural parameters turn out to model market specific allocations. This justifies the name *natural market allocation* for $\boldsymbol{\Theta}$, which may be viewed as the image by $\boldsymbol{\nabla}_{\psi}^{-1}$ of the market's expected returns. Taking as allocation matrix this natural market allocation, (15.18) represents a density of wealth associated to the support of market returns $\mathbf{W}$, as we have indeed:

$$p_{\psi}(\mathbf{W}; \boldsymbol{\Theta}) \propto \exp(\omega^F). \tag{15.22}$$

(15.22) us that the density of wealth is maximized for investments corresponding to the natural market allocation $\boldsymbol{\Theta}$, as the (unique) mode of exponential families occurs at their expectation parameters; furthermore, it happens that the natural market allocation is optimal from the information-theoretic standpoint (follows from Proposition 1 in [3], and (15.16) above).

Let us switch from the standpoint of the market to that of an investor. The famed St. Petersburg paradox tells us that this investor typically does not obey to the maximization of the expected value of reward, $\mathrm{E}_{\mathbf{W}\sim p_{\psi}}[\omega^F]$ [9]. In other words, as opposed to what (15.22) suggests, the investor would not follow maximum likelihood to fit his/her allocation. A more convenient framework, axiomatized by [18], considers that the investor maximizes instead the expected *utility* of reward, which boils down to maximizing in our case $\mathrm{E}_{\mathbf{W}\sim p_{\psi}}[\mathrm{U}(\omega^F)]$, where an *utility function* U models the investor's preferences in this framework. One usually requires that the first derivative of U be positive (non-satiation), and its second derivative be negative (risk-aversion). It can be shown that the expected utility equals the utility of the expected reward minus a real *risk premium* $\mathrm{P}_{\psi}(\mathbf{A}; \boldsymbol{\Theta})$:

$$\mathrm{E}_{\mathbf{W}\sim p_{\psi}}\left[\mathrm{U}(\omega^F)\right] = \mathrm{U}(\underbrace{\mathrm{E}_{\mathbf{W}\sim p_{\psi}}[\omega^F] - \mathrm{P}_{\psi}(\mathbf{A}; \boldsymbol{\Theta})}_{\mathrm{C}_{\psi}(\mathbf{A};\boldsymbol{\Theta})}). \tag{15.23}$$

It can further be shown that if the investor is risk-averse, the risk premium is strictly positive [9]. In this case, looking at the right-hand side of (15.23), we see that the risk premium acts like a penalty to the utility of the expected wealth. It represents a *shadow cost* to risk bearing in the context of market allocation, or, equivalently, the willingness of the investor to insure his/her portfolios.

There is one more remarkable thing about (15.23). While its left-hand side averages utilities over a potentially infinite number of markets, the right-hand side considers the utility of a *single case* which thus corresponds to a sure wealth equivalent to the left-hand side's numerous cases: it is called the *certainty equivalent* of the expected utility, $\mathrm{C}_{\psi}(\mathbf{A}; \boldsymbol{\Theta})$. What we have to do is derive, in the context of exponential families, the expressions of U, P_{ψ} and C_{ψ} in (15.23).

First, we adopt the usual landmarks that yield U [9, 23]. Consider the following Taylor approximations of the utility function around reward's expectation:

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

$$
\begin{aligned}
\mathrm{U}(\omega^F) \approx\ & \mathrm{U}(\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]) \\
& + (\omega^F - \mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]) \times \left.\frac{\partial}{\partial x}\mathrm{U}(x)\right|_{x=\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]} \\
& + \frac{(\omega^F - \mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F])^2}{2} \times \left.\frac{\partial^2}{\partial x^2}\mathrm{U}(x)\right|_{x=\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]},
\end{aligned}
\tag{15.24}
$$

$$
\begin{aligned}
\mathrm{U}(\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F] - \mathrm{P}_\psi(\mathbf{A};\boldsymbol{\Theta})) \approx\ & \mathrm{U}(\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]) \\
& - \mathrm{P}_\psi(\mathbf{A};\boldsymbol{\Theta}) \times \left.\frac{\partial}{\partial x}\mathrm{U}(x)\right|_{x=\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]}.
\end{aligned}
\tag{15.25}
$$

If we take expectations of (15.24) and (15.25), simplify taking into account the fact that $\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F - \mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]] = 0$, and match the resulting expressions using (15.23), we obtain the following approximate expression for the risk premium:

$$
\begin{aligned}
\mathrm{P}_\psi(\mathbf{A};\boldsymbol{\Theta}) \approx\ & \frac{1}{2}\mathrm{Var}_{\mathbf{W}\sim p_\psi}[\omega^F] \\
& \times \underbrace{\left\{ - \left.\frac{\partial^2}{\partial x^2}\mathrm{U}(x)\right|_{x=\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]} \left(\left.\frac{\partial}{\partial x}\mathrm{U}(x)\right|_{x=\mathrm{E}_{\mathbf{W}\sim p_\psi}[\omega^F]} \right)^{-1} \right\}}_{r(p_\psi)} \\
& .
\end{aligned}
\tag{15.26}
$$

Thus, approximation "in the small" of the risk premium makes it proportional to the variance of rewards *and* function $r(p_\psi)$, which is just, in the language of risk aversion, the Arrow–Pratt measure of *absolute risk aversion* [9, 23]. This expression for the risk premium is obviously not the one we shall use: its purpose is to shed light on the measure of absolute risk aversion, and derive the expression of U, as shown in the following Lemma.

Lemma 3. *$r(p_\psi) = k$, a constant matrix iff one of the following conditions holds true:*

$$
\begin{cases}
\mathrm{U}(x) = x & \text{if } \ k = 0 \\
\mathrm{U}(x) = -\exp(-ax) & \text{for some } \ a \in \mathbb{R}_* \text{ (otherwise)}
\end{cases}.
\tag{15.27}
$$

The proof of this Lemma is similar to the ones found in the literature (*e.g.* [9], Chap. 4). The framework of Lemma 3 is that of *constant absolute risk aversion* (CARA) [9], the framework on which we focus now, assuming that the investor is risk-averse. This implies $k \neq 0$ and $a > 0$; this constant a is called the *risk-aversion parameter*, and shall be implicit in some of our notations. We obtain the following expressions for C_ψ and P_ψ.

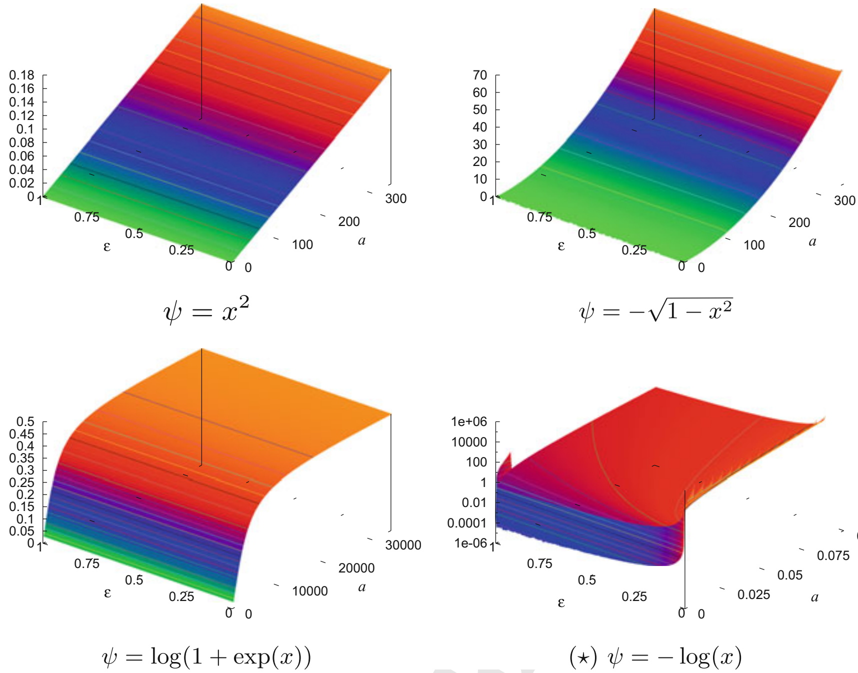

Fig. 15.1 Risk premia for various choices of generators, plotted as functions of the risk aversion parameter $a > 0$ and parameter $\varepsilon \in [0, 1]$ which modifies the natural market allocation (see text for the details of the model). Generators are indicated for each premium; see Table 15.1 for the associated Bregman matrix divergences. Symbol $(\star)$ indicates plots with logscale premium

Theorem 1. *Assume CARA and p_ψ defined as in (15.18). Then, the certainty equivalent and the risk premium associated to the portfolio are respectively:*

$$\mathrm{C}_\psi(\mathbf{A}; \boldsymbol{\Theta}) = \frac{1}{a}(\psi(\boldsymbol{\Theta}) - \psi(\boldsymbol{\Theta} - a\mathbf{A})), \quad (15.28)$$

$$\mathrm{P}_\psi(\mathbf{A}; \boldsymbol{\Theta}) = \frac{1}{a} D_\psi(\boldsymbol{\Theta} - a\mathbf{A} \| \boldsymbol{\Theta}). \quad (15.29)$$

Proof We first focus on the certainty equivalent. We have:

$$\begin{aligned} \mathrm{E}_{\mathbf{W} \sim p_\psi}[\mathrm{U}(\omega^F)] &= \int - \exp\left(\mathrm{Tr}\left(\mathbf{W}(\boldsymbol{\Theta} - a\mathbf{A})^\top\right) - \psi(\boldsymbol{\Theta})\right) b(\mathbf{W}) \mathrm{d}\mathbf{W} \\ &= - \exp\left(\psi(\boldsymbol{\Theta} - a\mathbf{A}) - \psi(\boldsymbol{\Theta})\right) \\ &\quad \times \underbrace{\int \exp\left(\mathrm{Tr}\left(\mathbf{W}(\boldsymbol{\Theta} - a\mathbf{A})^\top\right) - \psi(\boldsymbol{\Theta} - a\mathbf{A})\right) b(\mathbf{W}) \mathrm{d}\mathbf{W}}_{=1}. \end{aligned} \quad (15.30)$$

But we must also have from (15.23) and (15.27): $\mathrm{E}_{\mathbf{W}\sim p_\psi}[\mathrm{U}(\omega^F)] = -\exp\left(-a\mathrm{C}_\psi(\mathbf{A};\mathbf{W})\right)$. This identity together with (15.30) brings us expression (15.28). Now, for the risk premium, (15.23) brings:

$$\begin{aligned}
\mathrm{P}_\psi(\mathbf{A};\boldsymbol{\Theta}) &= \mathrm{E}_{\mathbf{W}\sim p_\psi}[\mathrm{U}(\omega^F)] - \mathrm{C}_\psi(\mathbf{A};\mathbf{W}) \\
&= \mathrm{Tr}\left(\mathbf{A}\nabla_\psi^\top(\boldsymbol{\Theta})\right) - \mathrm{C}_\psi(\mathbf{A};\mathbf{W}) \\
&= \frac{1}{a}\left(\psi(\boldsymbol{\Theta}-a\mathbf{A}) - \psi(\boldsymbol{\Theta}) + \mathrm{Tr}\left(a\mathbf{A}\nabla_\psi^\top(\boldsymbol{\Theta})\right)\right) \\
&= \frac{1}{a}D_\psi(\boldsymbol{\Theta}-a\mathbf{A}\|\boldsymbol{\Theta}), \qquad (15.31)
\end{aligned}$$

as claimed, where (15.31) uses the fact that $\mathrm{E}_{\mathbf{W}\sim p_\psi}[\mathrm{U}(\omega^F)] = \mathrm{E}_{\mathbf{W}\sim p_\psi}[\mathrm{Tr}\left(\mathbf{A}\mathbf{W}^\top\right)] = \mathrm{Tr}\left(\mathbf{A}\nabla_\psi^\top(\boldsymbol{\Theta})\right)$ from (15.21).

The following Lemma states among all that Theorem 1 is indeed a generalization of the mean-variance approach (proof straightforward).

Lemma 4. *The risk premium satisfies the following limit behaviors:*

$$\begin{aligned}
\lim_{a\to 0} \mathrm{P}_\psi(\boldsymbol{A};\boldsymbol{\Theta}) &= 0, \\
\lim_{\boldsymbol{A}\to_F \boldsymbol{Z}} \mathrm{P}_\psi(\boldsymbol{A};\boldsymbol{\Theta}) &= 0,
\end{aligned}$$

where $\to_F$ denotes the limit in Frobenius norm. Furthermore, when p_ψ is a multivariate Gaussian, the risk premium simplifies to the variance premium of the mean-variance model:

$$\mathrm{P}_\psi(\boldsymbol{A};\boldsymbol{\Theta}) = \frac{a}{2}\boldsymbol{diag}(\boldsymbol{A})^\top\boldsymbol{\Sigma}\boldsymbol{diag}(\boldsymbol{A}),$$

where $\boldsymbol{diag}(.)$ is the vector of the diagonal entries of the matrix.

One may use Lemma 4 as a sanity check for the risk premium, as the Lemma says that the risk premium tends to zero when risk aversion tends to zero, or when there is no allocation at all. Hereafter, we shall denote our generalized model as the *mean-divergence* model. Let us illustrate in a toy example the range of premia available, fixing the dimension to be $d = 1,000$. We let $\mathbf{A}$ and $\boldsymbol{\Theta}_\varepsilon$ be diagonal, where $\mathbf{A}$ denotes the uniform allocation ($\mathbf{A} = (1/d)\mathbf{I}$), and $\boldsymbol{\Theta}_\varepsilon$ depends on real $\varepsilon \in [0,1]$, with:

$$\theta_{ii} = \begin{cases} 1-\varepsilon & \text{if } \; i = 1, \\ \frac{\varepsilon}{d-1} & \text{otherwise} \end{cases}.$$

Thus, the natural market allocation shifts in between two extreme cases: the one in which the allocation emphasizes a single stock ($\varepsilon = 0$), and the one in which it is uniform on all but one stocks ($\varepsilon = 1$), admitting as intermediary setting the one in which the natural market allocation is uniform ($\varepsilon = (d-1)/d$). Risk premia are compared

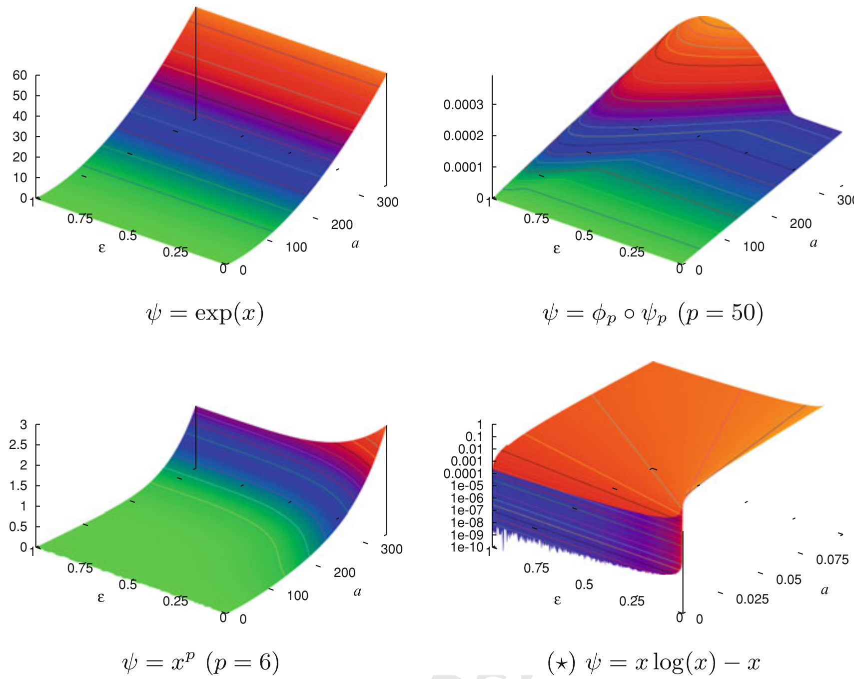

Fig. 15.2 More examples of risk premia. Conventions follow those of Fig. 15.1

against the mean-variance model's in which we let $\boldsymbol{\Sigma} = I$. The results are presented in Figs. 15.1 and 15.2. Notice that the mean-variance premium, which equals $a/(2d)$, displays the simplest behavior (a linear plot, see upper-left in Fig. 15.1).

15.4 On-line Learning in the Mean-Divergence Model

As previously studied by [14, 26] in the mean-variance model, our objective is now to track "efficient" portfolios at the market level, where a portfolio is all the more efficient as its associated risk premium (15.28) is reduced. Let us denote these portfolios *reference* portfolios, and the sequence of their allocation matrices as: $\mathbf{O}_0, \mathbf{O}_1, \ldots$. The natural market allocation may also shift over time, and we denote $\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_1, \ldots$ the sequence of natural parameter matrices of the market. Naturally, we could suppose that $\mathbf{O}_t = \boldsymbol{\Theta}_t, \forall t$, which would amount to tracking directly the natural market allocation, but this setting would be too restrictive because it may be easier to track some $\mathbf{O}_t$ close to $\boldsymbol{\Theta}_t$ but having specific properties that $\boldsymbol{\Theta}_t$ does not have (*e.g.* sparsity). Finally, we measure risk premia for references with the same risk aversion parameter a as for the investor's.

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

To adopt the same scale for allocation matrices, all shall be supposed to have r-norm upperbounded by ℓ, for some user-fixed $\ell > 0$ and $r > 0$. Assume for example $r = 1$: after division by ℓ, one can think such matrices as representing the way the investor scatters his/her wealth among the d stocks, leaving part of the wealth for a riskless investment if the trace is < 1. The algorithm we propose, simply named $\mathcal{A}$, uses ideas from Amari's natural gradient [1], to progress towards the minimization of the risk premium using a geometry induced by Bregman–Schatten p-divergence. To state this algorithm, we abbreviate the gradient (in $\mathbf{A}$) of the risk premium as:

$$\nabla_{\mathrm{P}_\psi}(\mathbf{A}; \boldsymbol{\Theta}) \doteq \nabla_\psi(\boldsymbol{\Theta}) - \nabla_\psi(\boldsymbol{\Theta} - a\mathbf{A})$$

(the risk aversion parameter a shall be implicit in the notation). Algorithm $\mathcal{A}$ initializes the following parameters: allocation matrix $\mathbf{A}_0 = \mathbf{Z}$, learning parameter $\eta_a > 0$, Bregman–Schatten parameter $q > 2$, and renormalization parameters $\ell > 0$ and $r > 0$; then, it proceeds through iterating what follows, for $t = 0, 1, \ldots, T-1$:

- (Premium dependent update) Upon receiving observed returns $\mathbf{W}_t$, compute $\boldsymbol{\Theta}_t$ using (15.21), and update portfolio allocation matrix to find the new *unnormalized* allocation matrix, $\mathbf{A}^u_{t+1}$:

$$\begin{aligned}\mathbf{A}^u_{t+1} &\leftarrow \nabla^{-1}_{\phi_q\circ\psi_q}(\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t) + \eta_a(\underbrace{s_t\mathbf{I} - \nabla_{\mathrm{P}_\psi}(\mathbf{A}_t; \boldsymbol{\Theta}_t)}_{\boldsymbol{\Delta}_t}))\\ &= \nabla_{\phi_p\circ\psi_p}(\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t) + \eta_a\boldsymbol{\Delta}_t)), \qquad (15.32)\end{aligned}$$

  $\forall t \geq 0$, with $s_t \geq 0$ picked to have $\boldsymbol{\Delta}_t$ positive definite. Lemma 1 implies the equality in (15.32).
- (Normalize) If $\left\|\mathbf{A}^u_{t+1}\right\|_r > \ell$ then $\mathbf{A}_{t+1} \leftarrow \left(\ell/\left\|\mathbf{A}^u_{t+1}\right\|_r\right)\mathbf{A}^u_{t+1}$, else $\mathbf{A}_{t+1} \leftarrow \mathbf{A}^u_{t+1}$.

We make the following assumption regarding market evolution: the matrix divergence or the risk premium is convex enough to exceed linear variations up to a small constant $\delta > 0$ (we let (**i**) denote this assumption):

$$\begin{aligned}\exists\delta > 0 : \forall t \geq 0,\ D_\psi(\boldsymbol{\Theta}_t - a\mathbf{O}_t\|\boldsymbol{\Theta}_t - a\mathbf{A}_t) &\geq \delta + s_t\mathrm{Tr}\left((\boldsymbol{\Theta}_t - a\mathbf{O}_t) - (\boldsymbol{\Theta}_t - a\mathbf{A}_t)\right)\\ &= \delta + as_t\mathrm{Tr}\left(\mathbf{A}_t - \mathbf{O}_t\right) \quad (\mathbf{i}).\end{aligned}$$

Let us denote

$$\mathbb{U} \doteq \{\boldsymbol{\Delta}_t, \forall t\} \cup \{\sum_{0\leq j<t}\boldsymbol{\Delta}_j, \forall t > 0\}.$$

This is the set of premium dependent updates, and all its elements are SPD matrices. We let $\lambda_* > 0$ denote the largest eigenvalue in the elements of $\mathbb{U}$, and $\rho_* \geq 1$ their largest eigenratio, where the eigenratio of a matrix is the ratio between its largest and smallest eigenvalues. We let $\mathbb{T}$ denote the set of indexes for which we perform

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

renormalization. Finally, we let

$$\nu_* \doteq \min\{1, \min_{t=1,2,\ldots,T} (\ell / \left\|\mathbf{A}_t^u\right\|_r)\} \quad (> 0),$$

which is 1 iff no renormalization has been performed. The following Theorem states that the total risk premium incurred by $\mathcal{A}$ basically deviates from that of the shifting reference by no more than two penalties: the first depends on the total shift of the reference, the second depends on the difference of the Schatten p-norms chosen for updating and renormalizing.

Theorem 2. *Pick*

$$0 < \eta_a < \frac{1}{\lambda_* d^{\frac{1}{2}-\frac{1}{q}} (1 + \nu_*^{-1} \rho_*)^{\frac{q}{2}-1}} \sqrt{\frac{2\delta}{a(q-1)}}.$$

Then, Algorithm $\mathcal{A}$ satisfies:

$$\begin{aligned} \sum_{t=0}^{T-1} \mathrm{P}_\psi(\mathbf{A}_t; \boldsymbol{\Theta}_t) \leq & \sum_{t=0}^{T-1} \mathrm{P}_\psi(\mathbf{O}_t; \boldsymbol{\Theta}_t) \\ & + \frac{1}{\eta_a} \left(b \left\|\mathbf{O}_T\right\|_r^2 + b^2 \ell \sum_{t=0}^{T-1} \left\|\mathbf{O}_{t+1} - \mathbf{O}_t\right\|_r + |\mathbb{T}| \ell^2 \left[d^{\frac{|q-r|}{qr}} - 1 \right]^2 \right). \end{aligned} \qquad (15.33)$$

Here, $b = 1$ iff $r \leq q$ and $b = d^{\frac{r-q}{qr}}$ otherwise.

Proof sketch: The proof makes an extensive use of two matrix inequalities that we state for symmetric matrices (but remain true in more general settings):

$$\left\|\mathbf{L}\right\|_\gamma d^{\frac{1}{\beta}-\frac{1}{\gamma}} \leq \left\|\mathbf{L}\right\|_\beta \leq \left\|\mathbf{L}\right\|_\gamma, \quad \forall \mathbf{L} \in \mathbb{R}^{d \times d}, \quad \forall \beta > \gamma > 0 \ ; \qquad (15.34)$$

$$\mathrm{Tr}\,(\mathbf{LN}) \leq \left\|\mathbf{L}\right\|_\beta \left\|\mathbf{N}\right\|_\gamma, \quad \forall \mathbf{L}, \mathbf{N} \in \mathbb{R}^{d \times d}, \forall \beta, \gamma \text{Hölder conjugates.} \qquad (15.35)$$

The former is a simple generalization of q-norm vector inequalities; the second is Hölder's matrix inequality. Following a general well-oiled technique [15], the proof consists in bounding a measure of progress to the shifting reference,

$$\delta_t \doteq D_{\phi_q \circ \psi_q}(\mathbf{O}_t \| \mathbf{A}_t) - D_{\phi_q \circ \psi_q}(\mathbf{O}_{t+1} \| \mathbf{A}_{t+1}). \qquad (15.36)$$

To take into consideration the possible renormalization, we split the progress into two parts, $\delta_{t,1}$, $\delta_{t,2}$, as follows:

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

$$\delta_t = \underbrace{D_{\phi_q\circ\psi_q}(\mathbf{O}_t\|\mathbf{A}_t) - D_{\phi_q\circ\psi_q}(\mathbf{O}_t\|\mathbf{A}^u_{t+1})}_{\delta_{t,1}}$$

$$+ \underbrace{D_{\phi_q\circ\psi_q}(\mathbf{O}_t\|\mathbf{A}^u_{t+1}) - D_{\phi_q\circ\psi_q}(\mathbf{O}_{t+1}\|\mathbf{A}_{t+1})}_{\delta_{t,2}}. \quad (15.37)$$

We now bound separately the two parts, starting with $\delta_{t,1}$. We have:

$$\delta_{t,1} = \eta_a \mathrm{Tr}\left((\mathbf{O}_t - \mathbf{A}_t)\boldsymbol{\Delta}_t\right) - D_{\phi_q\circ\psi_q}(\mathbf{A}_t\|\mathbf{A}^u_{t+1})$$

$$= \frac{\eta_a}{a}\underbrace{\mathrm{Tr}\left(((\boldsymbol{\Theta}_t - a\mathbf{A}_t) - (\boldsymbol{\Theta}_t - a\mathbf{O}_t))\left(\nabla_\psi(\boldsymbol{\Theta}_t - a\mathbf{A}_t) - \nabla_\psi(\boldsymbol{\Theta}_t)\right)\right)}_{\tau}$$

$$+\eta_a s_t \mathrm{Tr}\left(\mathbf{O}_t - \mathbf{A}_t\right) - D_{\phi_q\circ\psi_q}(\mathbf{A}_t\|\mathbf{A}^u_{t+1}). \quad (15.38)$$

The following Bregman triangle identity [19] holds true:

$$\tau = D_\psi(\boldsymbol{\Theta}_t - a\mathbf{O}_t\|\boldsymbol{\Theta}_t - a\mathbf{A}_t) + D_\psi(\boldsymbol{\Theta}_t - a\mathbf{A}_t\|\boldsymbol{\Theta}_t) - D_\psi(\boldsymbol{\Theta}_t - a\mathbf{O}_t\|\boldsymbol{\Theta}_t). \quad (15.39)$$

Plugging (15.39) in (15.38) and using assumption (**i**) yields:

$$\delta_{t,1} \geq \frac{\eta_a}{a}\left\{D_\psi(\boldsymbol{\Theta}_t - a\mathbf{A}_t\|\boldsymbol{\Theta}_t) - D_\psi(\boldsymbol{\Theta}_t - a\mathbf{O}_t\|\boldsymbol{\Theta}_t)\right\}$$

$$-D_{\phi_q\circ\psi_q}(\mathbf{A}_t\|\mathbf{A}^u_{t+1}) + \frac{\eta_a\delta}{a}. \quad (15.40)$$

Lemma 5. *The following bound holds for the divergence between successive updates:*

$$D_{\phi_q\circ\psi_q}(\mathbf{A}_t\|\mathbf{A}^u_{t+1}) \leq \frac{(q-1)\eta_a^2 d^{1-\frac{2}{q}}\left(1+\nu_*^{-1}\rho_*\right)^{q-2}\lambda_*^2}{2}. \quad (15.41)$$

Proof Plugging $\mathbf{L} \doteq \mathbf{A}_t$ and $\mathbf{N} \doteq \mathbf{A}^u_{t+1}$ in Lemma 1 (ii), and using (15.32), we get:

$$D_{\phi_q\circ\psi_q}(\mathbf{A}_t\|\mathbf{A}^u_{t+1}) = D_{\phi_p\circ\psi_p}(\underbrace{\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t) + \eta_a\boldsymbol{\Delta}_t}_{\mathbf{L}}\|\underbrace{\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)}_{\mathbf{N}}) \quad (15.42)$$

We now pick **L** and **N** as in (15.42), and use them in (15.13) (Lemma 2), along with the fact that $q > 2$ which ensures that ϕ_q is concave. There comes that there exists some $\alpha \in [0, 1]$ such that:

$$
\begin{aligned}
&(D_{\phi_q\circ\psi_q}(\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)+\eta_a\boldsymbol{\Delta}_t)||\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t))\\
&\leq \frac{\eta_a^2}{2}\left.\frac{\partial}{\partial x}\phi_q(x)\right|_{x=\psi_q\left(\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)\right)} \mathrm{Tr}\left(\boldsymbol{\Delta}_t^2\left.\frac{\partial^2}{\partial x^2}\psi_q(x)\right|_{x=\mathbf{U}_\alpha}\right)\\
&=\frac{(q-1)\eta_a^2}{2}\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)\right\|_q^{2-q}\mathrm{Tr}\left(\boldsymbol{\Delta}_t^2\,|\mathbf{U}_\alpha|^{q-2}\right),
\end{aligned}
\tag{15.43}
$$

with $\mathbf{U}_\alpha \doteq \nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)+\alpha\eta_a\boldsymbol{\Delta}_t$. We now use (15.35) with $\beta = q/(q-2)$ and $\gamma = q/2$, and we obtain $\mathrm{Tr}\left(\boldsymbol{\Delta}_t^2\,|\mathbf{U}_\alpha|^{q-2}\right) \leq \|\mathbf{U}_\alpha\|_q^{q-2}\,\|\boldsymbol{\Delta}_t\|_q^2$, which, using (15.43), yields the following bound on the divergence of $\tilde{\mathbf{A}}_{t+1}$ with respect to $\mathbf{A}_t$:

$$
\begin{aligned}
D_{\phi_q\circ\psi_q}(\mathbf{A}_t\|\tilde{\mathbf{A}}_{t+1}) &\leq \frac{(q-1)\eta_a^2}{2}\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)\right\|_q^{2-q}\|\mathbf{U}_\alpha\|_q^{q-2}\,\|\boldsymbol{\Delta}_t\|_q^2\\
&=\frac{(q-1)\eta_a^2}{2}\times\frac{\|\mathbf{U}_\alpha\|_q^{q-2}\,\|\boldsymbol{\Delta}_t\|_q^2}{\|\mathbf{A}_t\|_q^{-(q-2)^2}\left\|\mathbf{A}_t^{q-1}\right\|_q^{q-2}}.
\end{aligned}
\tag{15.44}
$$

We now work on $\|\mathbf{U}_\alpha\|_q$. Let υ denote an eigenvalue of $\mathbf{U}_\alpha$, and $\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t) = \mathbf{PDP}^\top$ the diagonalization of $\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)$. Bauer-Fike Theorem tells us that there exists an eigenvalue ϱ of $\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)$ such that:

$$
\begin{aligned}
|\upsilon-\varrho| &\leq \alpha\eta_a|\varrho|\,\|\mathbf{P}\|_F\left\|\mathbf{P}^\top\right\|_F\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)^{-1}\boldsymbol{\Delta}_t\right\|_F\\
&=\alpha\eta_a|\varrho|\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)^{-1}\boldsymbol{\Delta}_t\right\|_F,
\end{aligned}
\tag{15.45}
$$

because $\mathbf{P}$ is unitary. Denoting $\{\upsilon_i\}_{i=1}^d$ the (possibly multi-)set of non-negative eigenvalues of $\mathbf{U}_\alpha$, and $\{\varrho_i\}_{i=1}^d$ that of $\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)$, there comes from (15.45) that there exists $f:\{1,2,\ldots,d\}\to\{1,2,\ldots,d\}$ such that:

$$
\begin{aligned}
\|\mathbf{U}_\alpha\|_q \doteq \left(\sum_{i=1}^d \upsilon_i^q\right)^{\frac{1}{q}} &\leq \left(1+\alpha\eta_a\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)^{-1}\boldsymbol{\Delta}_t\right\|_F\right)\left(\sum_{i=1}^d \varrho_{f(i)}^q\right)^{\frac{1}{q}}\\
&\leq d^{\frac{1}{q}}\left(1+\eta_a\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)^{-1}\boldsymbol{\Delta}_t\right\|_F\right)\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)\right\|_\infty\\
&= d^{\frac{1}{q}}\left(1+\eta_a\left\|\nabla_{\phi_q\circ\psi_q}(\mathbf{A}_t)^{-1}\boldsymbol{\Delta}_t\right\|_F\right)\frac{\|\mathbf{A}_t\|_\infty^{q-1}}{\|\mathbf{A}_t\|_q^{q-2}}.
\end{aligned}
\tag{15.46}
$$

Putting (15.46) into (15.44) yields:

$$D_{\phi_q \circ \psi_q}(\mathbf{A}_t \| \tilde{\mathbf{A}}_{t+1}) \leq \frac{(q-1)\eta_a^2 d^{1-\frac{2}{q}} \left(1 + \eta_a \left\| \nabla_{\phi_q \circ \psi_q}(\mathbf{A}_t)^{-1} \boldsymbol{\Delta}_t \right\|_F \right)^{q-2} \|\boldsymbol{\Delta}_t\|_q^2}{2} \times \left(\frac{\|\mathbf{A}_t\|_\infty^{q-1}}{\left\| \mathbf{A}_t^{q-1} \right\|_q} \right)^{q-2} . \tag{15.47}$$

We now refine this bound in three steps. First, since $\|\mathbf{A}_t\|_\infty^{q-1} \leq \left\| \mathbf{A}_t^{q-1} \right\|_q$, the factor after the times is ≤ 1. Second, let us denote $\nu_* < \nu_t \leq 1$ the multiplicative factor by which we renormalize $\tilde{\mathbf{A}}_{t+1}$. Remarking that $\nabla_{\phi_q \circ \psi_q}(x\mathbf{L}) = |x| \nabla_{\phi_q \circ \psi_q}(\mathbf{L}), \forall x \in$
$\mathbb{R}_*$
and using Lemma 1, we obtain:

$$\begin{aligned}
\nabla_{\phi_q \circ \psi_q}(\mathbf{A}_t) &= \nabla_{\phi_q \circ \psi_q}\left(\nu_{t-1} \nabla_{\phi_p \circ \psi_p}(\nabla_{\phi_q \circ \psi_q}(\mathbf{A}_{t-1}) + \eta_a \boldsymbol{\Delta}_{t-1})\right) \\
&= \nu_{t-1} \nabla_{\phi_q \circ \psi_q}(\mathbf{A}_{t-1}) + \eta_a \nu_t \boldsymbol{\Delta}_{t-1} \\
&= \left(\prod_{j=0}^{t-1} \nu_j \right) \nabla_{\phi_q \circ \psi_q}(\mathbf{A}_0) + \eta_a \sum_{j=0}^{t-1} \left(\prod_{k=j}^{t-1} \nu_k \right) \boldsymbol{\Delta}_j \\
&\succeq \eta_a \nu_{t-1} \boldsymbol{\Delta}_{t-1} \succeq \mathbf{Z},
\end{aligned}$$

where $\mathbf{N} \succeq \mathbf{M}$ means $\mathbf{N} - \mathbf{M}$ is positive semi-definite. The rightmost inequality follows from the fact that the updates preserve the symmetric positive definiteness of $\mathbf{A}_{t+1}$. We get $\nabla_{\phi_q \circ \psi_q}(\mathbf{A}_t)^{-1} \preceq \eta_a^{-1} \pi_{t-1}^{-1} \boldsymbol{\Delta}_{t-1}^{-1}$, which, from Lemma 2 in [25], yields $\eta_a \left\| \nabla_{\phi_q \circ \psi_q}(\mathbf{A}_t)^{-1} \boldsymbol{\Delta}_t \right\|_F \leq \nu_{t-1}^{-1} \left\| \boldsymbol{\Delta}_{t-1}^{-1} \boldsymbol{\Delta}_t \right\|_F \leq \nu_{t-1}^{-1} \rho_* \leq \nu_*^{-1} \rho_*$. Third and last, $\|\boldsymbol{\Delta}_t\|_q \leq \lambda_*$. Plugging these three refinements in (15.47) yields the statement of the Lemma.

Armed with the statement of Lemma 5 and the upperbound on η_a, we can refine (15.40) and obtain our lowerbound on $\delta_{t,1}$ as:

$$\delta_{t,1} \geq \frac{\eta_a}{a} \left\{ D_\psi(\boldsymbol{\Theta}_t - a\mathbf{A}_t \| \boldsymbol{\Theta}_t) - D_\psi(\boldsymbol{\Theta}_t - a\mathbf{O}_t \| \boldsymbol{\Theta}_t) \right\} . \tag{15.48}$$

We now work on $\delta_{t,2}$. We distinguish two cases:

Case 1 $\left\| \mathbf{A}_{t+1}^u \right\|_r \leq \ell$ (we do not perform renormalization). In this case, $\mathbf{A}_{t+1} = \mathbf{A}_{t+1}^u$. Using (15.35) with $\beta = q, \gamma = q/(q-1)$ which brings

$$\mathrm{Tr}\left(\mathbf{L} \nabla_{\phi_q \circ \psi_q}(\mathbf{A}_{t+1})\right) \leq \|\mathbf{L}\|_q \|\mathbf{A}_{t+1}\|_q ,$$

we easily obtain the lowerbound:

$$
\begin{aligned}
& D_{\phi_q \circ \psi_q}(\mathbf{O}_t \| \mathbf{A}^u_{t+1}) - D_{\phi_q \circ \psi_q}(\mathbf{O}_{t+1} \| \mathbf{A}_{t+1}) \\
& \geq \frac{1}{2} \|\mathbf{O}_t\|_q^2 - \frac{1}{2} \|\mathbf{O}_{t+1}\|_q^2 - \|\mathbf{O}_{t+1} - \mathbf{O}_t\|_q \|\mathbf{A}_{t+1}\|_q . \qquad (15.49)
\end{aligned}
$$

Case 2 $\left\|\mathbf{A}^u_{t+1}\right\|_r > \ell$ (we perform renormalization). Because the reference matrix satisfies $\|\mathbf{O}_t\|_r \leq \ell$, renormalization implies $\|\mathbf{O}_t\|_r \leq \|\mathbf{A}_{t+1}\|_r$. This inequality, together with (15.34), brings:

$$\|\mathbf{O}_t\|_q \leq \|\mathbf{A}_{t+1}\|_q \, d^{\frac{|q-r|}{qr}} .$$

Using the shorthands:

$$
\begin{aligned}
u_{t+1} &\doteq \frac{\ell}{\left\|\mathbf{A}^u_{t+1}\right\|_r} \quad (\in (0,1)), \\
v &\doteq 2d^{\frac{|q-r|}{qr}} \quad (\geq 2), \\
g(x,y) &\doteq \frac{(1-x)(y-x)}{x^2},
\end{aligned}
$$

and one more application of (15.35) as in Case 1, we obtain:

$$
\begin{aligned}
& D_{\phi_q \circ \psi_q}(\mathbf{O}_t \| \mathbf{A}^u_{t+1}) - D_{\phi_q \circ \psi_q}(\mathbf{O}_{t+1} \| \mathbf{A}_{t+1}) \\
& \geq \frac{1}{2} \|\mathbf{O}_t\|_q^2 - \frac{1}{2} \|\mathbf{O}_{t+1}\|_q^2 \\
& + \frac{v-1}{2} g\left(u_{t+1}, \frac{1}{v-1}\right) \|\mathbf{A}_{t+1}\|_q^2 - \|\mathbf{O}_{t+1} - \mathbf{O}_t\|_q \|\mathbf{A}_{t+1}\|_q . \qquad (15.50)
\end{aligned}
$$

We are now in a position to bring (15.49) and (15.50) altogether: summing for $t = 0, 1, \ldots, T-1$ (15.37) using (15.48) and (15.50), we get:

$$
\begin{aligned}
D_{\phi_q \circ \psi_q}(\mathbf{O}_0 \| \mathbf{A}_0) - D_{\phi_q \circ \psi_q}(\mathbf{O}_T \| \mathbf{A}_T) &= \sum_{t=0}^{T-1} \delta_t \\
& \geq \eta_a \sum_{t=0}^{T-1} \mathrm{P}_\psi(\mathbf{A}_t; \boldsymbol{\Theta}_t) - \eta_a \sum_{t=0}^{T-1} \mathrm{P}_\psi(\mathbf{O}_t; \boldsymbol{\Theta}_t) \\
& + \frac{1}{2} \|\mathbf{O}_0\|_q^2 - \frac{1}{2} \|\mathbf{O}_T\|_q^2 - \sum_{t=0}^{T-1} \|\mathbf{O}_{t+1} - \mathbf{O}_t\|_q \|\mathbf{A}_{t+1}\|_q \\
& + \frac{v-1}{2} \sum_{t \in \mathbb{T}} g\left(u_t, \frac{1}{v-1}\right) \|\mathbf{A}_t\|_q^2, \qquad (15.51)
\end{aligned}
$$

where we recall that $\mathbb{T}$ contains the indexes of renormalization updates. Because $g(x,y) \geq -(1-y)^2/(4y)$, the following lowerbound holds:

$$g\left(u_t, \frac{1}{\upsilon - 1}\right) \geq -\frac{\upsilon - 2}{4}, \forall t \in \mathbb{T}.$$

There remains to plug this bound into (15.51) and simplify a bit further to obtain the statement of the Theorem. □

The bound in Theorem 33 shows that the sum of premia of algorithm $\mathcal{A}$ is no larger than the sum of premia of *any* sequence of shifting references *plus* two penalties: the first depends on the sequence of references; the second (the rightmost term in (15.33)) is structural as it is zero when $q = r$. Both penalties are proportional to $\sqrt{a}$: they are thus *sublinear* on the risk aversion parameter. This is interesting, as one can show that the risk premium is always *superlinear* in a, with the exception of Markowitz' mean-variance model for which it is linear (see Fig. 15.1). Hence, the effects of risk aversion in the penalty are much smaller than in the premia. Finally, we can note that if small premia are achieved by reference allocations with sparse eigenspectra and that do not shift too much over periods, then the premia of $\mathcal{A}$ shall be small as well.

15.5 Experiments on Learning in the Mean-Divergence Model

We have made a toy experiment of $\mathcal{A}$ over the $d = 324$ stocks which belonged to the S&P 500 over the periods ranging from 01/08/1998 to 11/12/2009 (1 period = 1 week, $T = 618$). Our objective in performing these few experiments is not to show whether $\mathcal{A}$ competes with famed experimental approaches like [5]. Clearly, we have not tuned the parameters of $\mathcal{A}$ to obtain the best-looking results in Fig. 15.3. Our objective is rather to display on a real market and over a sufficiently large number of iterations (i) whether the mean-divergence model can be useful to spot insightful market events, and (ii) wether simple on-line learning approaches, grounded on a solid theory, can effectively track reduced risk portfolios, obtain reasonably large certainty equivalents, and thus suggest that the mean-divergence model may be a valuable starting point for much more sophisticated approaches [5]. Figure 15.3 displays comparisons between $\mathcal{A}$ and the Uniform Cost Rebalanced Portfolio ($\mathcal{UCRP}$), which consists in equally scattering wealth among stocks. The Figure also displays the Kullback–Leibler divergence between two successive portfolios for $\mathcal{A}$ (this would be zero for $\mathcal{UCRP}$): the higher the divergence, the higher the differences between successive portfolios selected by $\mathcal{A}$. We see from the pictures that $\mathcal{A}$ manages significant variations of its portfolio through iterations (divergence almost always > 0.05), yet it does turn like a weather vane through market periods (divergence almost always < 0.3). The fact that market accidents make the divergence peak, like during the subprime crisis ($T > 500$), indicate that the algorithm significantly reallocates its portfolio during such events. As shown in the Figure, this is achieved with certain success compared to the $\mathcal{UCRP}$. Figure 15.4 displays risk premia for $\mathcal{A}$ when shifting from Markowitz' premium to that induced by the logdet divergence, a premium which displays by far the steepest variations among premia in Figs. 15.1 and 15.2. Figure 15.4

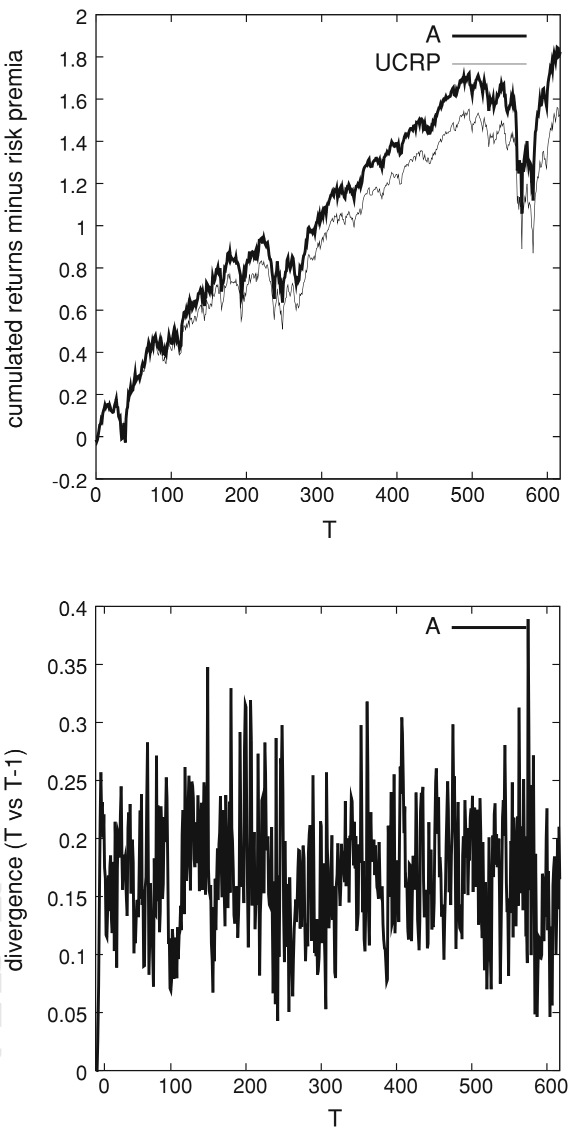

Fig. 15.3 *Up* Comparison of cumulated returns minus premia (certainty equivalents) for $\mathcal{A}$ (*bold lines*) versus the Uniform Cost Rebalanced Portfolio ($\mathcal{UCRP}$, *thin lines*). Parameters for the algorithms are: $a = 100$, $r = \ell = 1, q = 2.1, \eta = 100$, premium divergence $=$ Mahalanobis. *Down* Kullback–Leibler divergence between two successive portfolios for $\mathcal{A}$

displays the relevance of the generalized mean-divergence model. Changing the premium generator may indeed yield to dramatic peaks of premia that can alert the investor on significant events at the market scale, like in Fig. 15.4, for which the tallest peaks appear during the subprime crisis.

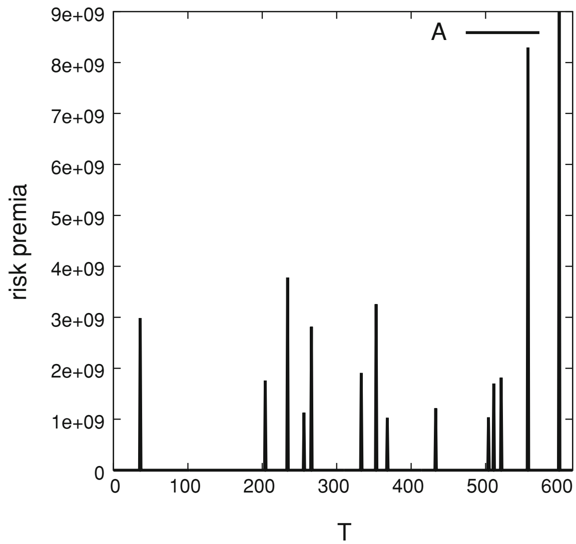

Fig. 15.4 Premia for $\mathcal{A}$, with $a = 100, r = \ell = 1, q = 4$, $\eta = 100$, premium divergence = logdet (Table 15.1). See text for details

15.6 Discussion

In this section, our objective is twofold. first, we drill down into the properties of our divergences (15.2), and compare them to the properties of other matrix divergences based on Bregman divergences published elsewhere. Second, we exploit these properties to refine our analysis on the risk premium of our mean-divergence model. Thus, for our first goal, the matrix arguments of the divergences are not assumed to be symmetric anymore.

Reference [13] have previously defined a particular case of matrix-based divergence, which corresponds to computing the usual p-norm vector divergence between spec $(\mathbf{L})$ and spec $(\mathbf{N})$. It is not hard to check that this corresponds to a particular case of Bregman–Schatten p-divergences in the case where one assumes that $\mathbf{L}$ and $\mathbf{N}$ share the same transition matrix. The qualitative gap between the definitions is significant: in the case of a general Bregman matrix divergences, such an assumption would make the divergence *separable*, that is, summing coordinate-wise divergences [11]. This is what the following Theorem shows. We adapt notation (15.4) to vectors and define $\tilde{\boldsymbol{u}}$ the vector with coordinates $\nabla_\psi(u_i)$. We also make use of the Hadamard product $\cdot$ previously used in Table 15.1.

Theorem 3. *Assume diagonalizable squared matrices* $\mathbf{L}$ *and* $\mathbf{N}$, *with their diagonalizations respectively denoted:*

$$\boldsymbol{L} = \boldsymbol{P}_{\mathrm{L}} \boldsymbol{D}_{\mathrm{L}} \boldsymbol{P}_{\mathrm{L}}^{-1},$$
$$\boldsymbol{N} = \boldsymbol{P}_{\mathrm{N}} \boldsymbol{D}_{\mathrm{N}} \boldsymbol{P}_{\mathrm{N}}^{-1}.$$

Denote the (non necessarily distinct) eigenvalues of $\boldsymbol{L}$ *(resp.* $\boldsymbol{N}$*) as:* $\lambda_1, \lambda_2, \ldots, \lambda_d$ *(resp.* $\nu_1, \nu_2, \ldots, \nu_d$*), and the corresponding eigenvectors as:* $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_d$ *(resp.* $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_d$*). Finally, let* $\boldsymbol{\lambda} \doteq \boldsymbol{diag}(\boldsymbol{D}_{\mathbf{L}})$, $\boldsymbol{\nu} \doteq \boldsymbol{diag}(\boldsymbol{D}_{\mathbf{N}})$ *and*

$$\boldsymbol{\Pi}_{\mathbf{X},\mathbf{Y}} \doteq \boldsymbol{P}_{\mathbf{X}}^\top \boldsymbol{P}_{\mathbf{Y}}, \forall \boldsymbol{X}, \boldsymbol{Y} \in \{\boldsymbol{L}, \boldsymbol{N}\},$$

$$\boldsymbol{H}_{\mathbf{X},\mathbf{Y}} \doteq \boldsymbol{\Pi}_{\mathbf{X},\mathbf{Y}}^{-1} \cdot \boldsymbol{\Pi}_{\mathbf{X},\mathbf{Y}}^\top.$$

Then any Bregman matrix divergence can be written as:

$$D_\psi(\boldsymbol{L}||\boldsymbol{N}) = \sum_{i=1}^{d} D_\psi(\lambda_i||\nu_i) + \boldsymbol{\lambda}^\top(\boldsymbol{I} - \boldsymbol{H}_{\mathbf{N},\mathbf{L}})\tilde{\boldsymbol{\nu}} + \boldsymbol{\nu}^\top(\boldsymbol{H}_{\mathbf{N},\mathbf{N}} - \boldsymbol{I})\tilde{\boldsymbol{\nu}}. \tag{15.52}$$

If, in addition, $\boldsymbol{N}$ *is symmetric, (15.52) becomes:*

$$D_\psi(\boldsymbol{L}||\boldsymbol{N}) = \sum_{i=1}^{d} D_\psi(\lambda_i||\nu_i) + \boldsymbol{\lambda}^\top(\boldsymbol{I} - \boldsymbol{H}_{\mathbf{N},\mathbf{L}})\tilde{\boldsymbol{\nu}}, \tag{15.53}$$

If, in addition, $\boldsymbol{L}$ *is symmetric, (15.53) holds for some doubly-stochastic* $\boldsymbol{H}_{\mathbf{N},\mathbf{L}}$*. If, in addition,* $\boldsymbol{L}$ *and* $\boldsymbol{N}$ *share the same transition matrices (*$\boldsymbol{P}_{\mathbf{L}} = \boldsymbol{P}_{\mathbf{N}}$*), (15.53) becomes:*

$$D_\psi(\boldsymbol{L}||\boldsymbol{N}) = \sum_{i=1}^{d} D_\psi(\lambda_i||\nu_i). \tag{15.54}$$

Proof Calling to (15.1) and using the general definition of (15.2), we get:

$$D_\psi(\mathbf{L}||\mathbf{N}) = \mathrm{Tr}\left(\sum_{k\geq 0} t_{\psi,k}\mathbf{L}^k\right) - \mathrm{Tr}\left(\sum_{k\geq 0} t_{\psi,k}\mathbf{N}^k\right) - \mathrm{Tr}\left(\sum_{k\geq 0} t_{\nabla_\psi,k}(\mathbf{L}-\mathbf{N})(\mathbf{N}^\top)^k\right).$$

Introducing the diagonalization, we obtain:

$$\begin{aligned} D_\psi(\mathbf{L}||\mathbf{N}) &= \mathrm{Tr}\left(\mathbf{P}_{\mathbf{L}}\left(\sum_{k\geq 0} t_{\psi,k}\mathbf{D}_{\mathbf{L}}^k\right)\mathbf{P}_{\mathbf{L}}^{-1}\right) - \mathrm{Tr}\left(\mathbf{P}_{\mathbf{N}}\left(\sum_{k\geq 0} t_{\psi,k}\mathbf{D}_{\mathbf{N}}^k\right)\mathbf{P}_{\mathbf{N}}^{-1}\right) \\ &\quad \underbrace{-\,\mathrm{Tr}\left(\mathbf{L}\sum_{k\geq 0} t_{\nabla_\psi,k}(\mathbf{N}^\top)^k\right)}_{a} + \underbrace{\mathrm{Tr}\left(\mathbf{N}\sum_{k\geq 0} t_{\nabla_\psi,k}(\mathbf{N}^\top)^k\right)}_{b} \\ &= \sum_{i=1}^{d}\psi(\lambda_i) - \sum_{i=1}^{d}\psi(\nu_i) - a + b. \end{aligned} \tag{15.55}$$

Now, using the cyclic invariance of the trace and the definition of $\mathbf{H}_{\mathbf{N},\mathbf{L}}$, we get:

$$\begin{aligned} a &= \mathrm{Tr}\left(\mathbf{P}_{\mathbf{L}}\mathbf{D}_{\mathbf{L}}\mathbf{P}_{\mathbf{L}}^{-1}(\mathbf{P}_{\mathbf{N}}^{-1})^{\top}\left(\sum_{k\geq 0} t_{\nabla_{\psi},k}\mathbf{D}_{\mathbf{N}}^{k}\right)\mathbf{P}_{\mathbf{N}}^{\top}\right) \\ &= \mathrm{Tr}\left(\mathbf{D}_{\mathbf{L}}\boldsymbol{\Pi}_{\mathbf{N},\mathbf{L}}^{-1}\left(\sum_{k\geq 0} t_{\nabla_{\psi},k}\mathbf{D}_{\mathbf{N}}^{k}\right)\boldsymbol{\Pi}_{\mathbf{N},\mathbf{L}}\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i(\pi^{-1})_{ij}\tilde{\nu}_j\pi_{ji} = \boldsymbol{\lambda}^{\top}\mathbf{H}_{\mathbf{N},\mathbf{L}}\tilde{\boldsymbol{\nu}}. \end{aligned} \tag{15.56}$$

Here, we have made use of π_{ij}, the general term of $\boldsymbol{\Pi}_{\mathbf{N},\mathbf{L}}$, and $(\pi^{-1})_{ij}$, the general term of $\boldsymbol{\Pi}_{\mathbf{N},\mathbf{L}}^{-1} = \mathbf{P}_{\mathbf{L}}^{-1}(\mathbf{P}_{\mathbf{N}}^{\top})^{-1} = \mathbf{P}_{\mathbf{L}}^{-1}(\mathbf{P}_{\mathbf{N}}^{-1})^{\top}$. Using the same path, we obtain:

$$\begin{aligned} b &= \mathrm{Tr}\left(\mathbf{P}_{\mathbf{N}}\mathbf{D}_{\mathbf{N}}\mathbf{P}_{\mathbf{N}}^{-1}(\mathbf{P}_{\mathbf{N}}^{-1})^{\top}\left(\sum_{k\geq 0} t_{\nabla_{\psi},k}\mathbf{D}_{\mathbf{N}}^{k}\right)\mathbf{P}_{\mathbf{N}}^{\top}\right) \\ &= \mathrm{Tr}\left(\mathbf{D}_{\mathbf{N}}\boldsymbol{\Pi}_{\mathbf{N},\mathbf{N}}^{-1}\left(\sum_{k\geq 0} t_{\nabla_{\psi},k}\mathbf{D}_{\mathbf{N}}^{k}\right)\boldsymbol{\Pi}_{\mathbf{N},\mathbf{N}}\right) = \boldsymbol{\nu}^{\top}\mathbf{H}_{\mathbf{N},\mathbf{N}}\tilde{\boldsymbol{\nu}}. \end{aligned} \tag{15.57}$$

Plugging (15.56) and (15.57) in (15.55) yields:

$$\begin{aligned} D_{\psi}(\mathbf{L}||\mathbf{N}) &= \sum_{i=1}^{d}\psi(\lambda_i) - \sum_{i=1}^{d}\psi(\nu_i) + \boldsymbol{\nu}^{\top}\mathbf{H}_{\mathbf{N},\mathbf{N}}\tilde{\boldsymbol{\nu}} - \boldsymbol{\lambda}^{\top}\mathbf{H}_{\mathbf{N},\mathbf{L}}\tilde{\boldsymbol{\nu}} \\ &= \sum_{i=1}^{d} D_{\psi}(\lambda_i||\nu_i) + \boldsymbol{\lambda}^{\top}\mathbf{I}\tilde{\boldsymbol{\nu}} - \boldsymbol{\nu}^{\top}\mathbf{I}\tilde{\boldsymbol{\nu}} + \boldsymbol{\nu}^{\top}\mathbf{H}_{\mathbf{N},\mathbf{N}}\tilde{\boldsymbol{\nu}} - \boldsymbol{\lambda}^{\top}\mathbf{H}_{\mathbf{N},\mathbf{L}}\tilde{\boldsymbol{\nu}} \\ &= \sum_{i=1}^{d} D_{\psi}(\lambda_i||\nu_i) + \boldsymbol{\lambda}^{\top}(\mathbf{I} - \mathbf{H}_{\mathbf{N},\mathbf{L}})\tilde{\boldsymbol{\nu}} + \boldsymbol{\nu}^{\top}(\mathbf{H}_{\mathbf{N},\mathbf{N}} - \mathbf{I})\tilde{\boldsymbol{\nu}}, \end{aligned} \tag{15.58}$$

as claimed. When $\mathbf{N}$ is symmetric, we easily get $\mathbf{H}_{\mathbf{N},\mathbf{L}} = \mathbf{I}$, and we obtain (15.54). If, in addition, $\mathbf{N}$ is symmetric, both transition matrices $\mathbf{P}_{\mathbf{L}}$ and $\mathbf{P}_{\mathbf{N}}$ are unitary. In this case, $m_{ij} = \boldsymbol{l}_i^{\top}\boldsymbol{n}_j = (m^{-1})_{ji}$, and so $q_{ij} = (\boldsymbol{l}_i^{\top}\boldsymbol{n}_j) = \cos^2(\boldsymbol{l}_i, \boldsymbol{n}_j) = q_{ji} \geq 0$, which yields $\sum_{j=1}^{d} q_{ij} = \sum_{j=1}^{d}\cos^2(\boldsymbol{l}_i, \boldsymbol{n}_j) = 1$, and so $\mathbf{H}_{\mathbf{N},\mathbf{L}}$ is doubly stochastic. To finish up, when, in addition, $\mathbf{L}$ and $\mathbf{N}$ share the same transition matrices, we immediately get $\mathbf{H}_{\mathbf{N},\mathbf{L}} = \mathbf{I}$, and we obtain (15.54). □

Hence, $D_{\psi}(\mathbf{L}||\mathbf{N})$ can be written in the form of a *separable* term plus two penalties: $D_{\psi}(\mathbf{L}||\mathbf{N}) = \sum_{i=1}^{d} D_{\psi}(\lambda_i||\nu_i) + p_1 + p_2$, where $p_1 \doteq \boldsymbol{\nu}^{\top}(\mathbf{H}_{\mathbf{N},\mathbf{N}} - \mathbf{I})\tilde{\boldsymbol{\nu}}$ is zero when

$\mathbf{N}$ is symmetric, and $p_2 \doteq \boldsymbol{\lambda}^\top(\mathbf{I} - \mathbf{H}_{\mathbf{N},\mathbf{L}})\tilde{\boldsymbol{\nu}}$ is zero when $\mathbf{L}$ and $\mathbf{N}$ are symmetric and share the same transition matrices.

The definition of Bregman matrix divergences makes quite a large consensus, yet some variations do exist. For example, [12, 16] use a very particular composition of two functions, $\phi \circ \psi$, in which ϕ is actually the divergence generator and ψ lists the eigenvalues of the matrix. In this case, (15.52) would be replaced by (writing for short $\mathbf{H}$ instead of $\mathbf{H}_{\mathbf{N},\mathbf{L}}$ hereafter):

$$D_\psi(\mathbf{L}||\mathbf{N}) = \mathrm{Tr}\left(\mathbf{D}_\psi \mathbf{H}\right), \tag{15.59}$$

where $\mathbf{D}_\psi$ is the divergence matrix whose general (i, j) term is $D_\psi(\lambda_i||\nu_j)$. Let us compare (15.59) to (15.53) when both arguments are symmetric matrices — which is the case for our finance application —, which can be abbreviated as:

$$D_\psi(\mathbf{L}||\mathbf{N}) = \mathrm{Tr}\left(\mathbf{D}_\psi\right) + \boldsymbol{\lambda}^\top(\mathbf{I} - \mathbf{H})\tilde{\boldsymbol{\nu}}. \tag{15.60}$$

We see that (15.60) clearly separates the divergence term ($\mathbf{D}_\psi$) from an *interaction* term, which depends on both the eigenvectors (transition matrices) and eigenvalues: $\boldsymbol{\lambda}^\top(\mathbf{I} - \mathbf{H})\tilde{\boldsymbol{\nu}}$. If we move back to our generalization of the mean-variance model, we have $\mathbf{L} = \boldsymbol{\Theta} - a\mathbf{A}$ and $\mathbf{N} = \boldsymbol{\Theta}$ ($\boldsymbol{\Theta}$ and $\mathbf{A}$ are symmetric). Adding term $a\mathbf{A}$ to $\boldsymbol{\Theta}$ possibly changes the transition matrix compared to $\boldsymbol{\Theta}$, and so produces a non-null interaction term between stocks. Furthermore, as the allocation $\mathbf{A}$ gets different from the natural market allocation $\boldsymbol{\Theta}$, and as the risk aversion a increases, so tends to do the magnitude of the interaction term. To study further its magnitude, let us define:

$$\varsigma \doteq \|\mathbf{I} - \mathbf{H}\|_F. \tag{15.61}$$

We analyze ς when the risk term $a\mathbf{A}$ remains sufficiently small, which amounts to assuming reduced risk premia as well. For this objective, recalling that both $\boldsymbol{\Theta}$ and $\mathbf{A}$ are SPD, we denote their eigensystems as follows:

$$\boldsymbol{\Theta}\mathbf{T} = \mathbf{T}\mathbf{D}, \tag{15.62}$$

$$(\boldsymbol{\Theta} - a\mathbf{A})\mathbf{V} = \mathbf{V}\mathbf{D}', \tag{15.63}$$

where the columns of $\mathbf{T}$, (resp. $\mathbf{V}$) are the eigenvectors and the diagonal elements of diagonal matrix $\mathbf{D}$ (resp. $\mathbf{D}'$) are the corresponding eigenvalues. The geometric multiplicity of eigenvalue d_{ii} is denoted $\mathfrak{g}(d_{ii})$. We say that the *first-order shift setting* holds when the second-order variations in the eigensystem of $\boldsymbol{\Theta}$ due to the shift $a\mathbf{A}$ are negligible, that is, when:

$$a\mathbf{A}(\mathbf{V} - \mathbf{T}) \approx (\mathbf{V} - \mathbf{T})(\mathbf{D}' - \mathbf{D}) \approx (\mathbf{V} - \mathbf{T})^\top(\mathbf{V} - \mathbf{T}) \approx \mathbf{Z}. \tag{15.64}$$

Lemma 6. *Under the first-order shift setting, the following holds true on the eigensystems (15.62) and (15.63):*

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

$$\boldsymbol{diag}(\boldsymbol{D}' - \boldsymbol{D}) = -a\boldsymbol{diag}(\boldsymbol{T}^\top \boldsymbol{A}\boldsymbol{T}) \quad (15.65)$$
$$\boldsymbol{V} - \boldsymbol{T} = \boldsymbol{T}\boldsymbol{B}, \quad (15.66)$$

with $\boldsymbol{B}$ a matrix whose general term b_{ij} satisfies:

$$b_{ij} = \begin{cases} 0 & \textit{if}(\mathfrak{g}(d_{ii}) > 1) \vee (\mathfrak{g}(d_{jj}) > 1) \vee (i = j) \\ \frac{a\boldsymbol{t}_i^\top \boldsymbol{A}\boldsymbol{t}_j}{d_{ii} - d_{jj}} & \textit{otherwise} \end{cases} \quad (15.67)$$

Here, $\boldsymbol{t}_i$ is the eigenvector in column i of $\boldsymbol{T}$, and d_{ii} its eigenvalue.

Proof sketch: The proof stems from standard linear algebra arguments [24]. We distinguish two cases:

Case 1 all eigenvalues have geometric multiplicity $\mathfrak{g}(.) = 1$. Denote for short $\mathbf{V} = \mathbf{T} + \boldsymbol{\Delta}$ and $\mathbf{D}' = \mathbf{D} + \boldsymbol{\Lambda}$. We have:

$$(\boldsymbol{\Theta} - a\mathbf{A})\mathbf{V} = \mathbf{V}\mathbf{D}'$$
$$\Leftrightarrow \boldsymbol{\Theta}\boldsymbol{\Delta} - a\mathbf{A}\mathbf{T} - a\mathbf{A}\boldsymbol{\Delta} = \mathbf{T}\boldsymbol{\Lambda} + \boldsymbol{\Delta}\mathbf{D} + \boldsymbol{\Delta}\boldsymbol{\Lambda}$$
$$\Leftrightarrow \boldsymbol{\Theta}\boldsymbol{\Delta} - a\mathbf{A}\mathbf{T} = \mathbf{T}\boldsymbol{\Lambda} + \boldsymbol{\Delta}\mathbf{D},$$

where we have used the fact that $\boldsymbol{\Theta}\mathbf{T} = \mathbf{T}\mathbf{D}$, $a\mathbf{A}\boldsymbol{\Delta} \approx \mathbf{Z}$ and $\boldsymbol{\Delta}\boldsymbol{\Lambda} \approx \mathbf{Z}$. Because of the assumption of the Lemma, the columns of $\mathbf{T}$ induce an orthonormal basis of $\mathbb{R}^d$, so that we can search for the coordinates of the columns of $\boldsymbol{\Delta}$ in this basis, which means finding $\mathbf{B}$ with:

$$\boldsymbol{\Delta} = \mathbf{T}\mathbf{B}. \quad (15.68)$$

Column i in $\mathbf{B}$ denotes the coordinates of column i in $\boldsymbol{\Delta}$ according to the eigenvectors in the columns of $\mathbf{T}$. We get

$$\boldsymbol{\Theta}\mathbf{T}\mathbf{B} - a\mathbf{A}\mathbf{T} = \mathbf{T}\boldsymbol{\Lambda} + \mathbf{T}\mathbf{B}\mathbf{D}$$
$$\Leftrightarrow \mathbf{T}\mathbf{D}\mathbf{B} - a\mathbf{A}\mathbf{T} = \mathbf{T}\boldsymbol{\Lambda} + \mathbf{T}\mathbf{B}\mathbf{D}$$
$$\Leftrightarrow \mathbf{T}^\top\mathbf{T}\mathbf{D}\mathbf{B} - a\mathbf{T}^\top\mathbf{A}\mathbf{T} = \mathbf{T}^\top\mathbf{T}\boldsymbol{\Lambda} + \mathbf{T}^\top\mathbf{T}\mathbf{B}\mathbf{D}$$
$$\Leftrightarrow \mathbf{D}\mathbf{B} - a\mathbf{T}^\top\mathbf{A}\mathbf{T} = \boldsymbol{\Lambda} + \mathbf{B}\mathbf{D},$$

i.e.:

$$\boldsymbol{\Lambda} = \mathbf{D}\mathbf{B} - \mathbf{B}\mathbf{D} - a\mathbf{T}^\top\mathbf{A}\mathbf{T}. \quad (15.69)$$

We have used the following facts: $\boldsymbol{\Theta}\mathbf{T} = \mathbf{T}\mathbf{D}$ and $\mathbf{T}^\top\mathbf{T} = \mathbf{I}$ ($\mathbf{T}^\top = \mathbf{T}^{-1}$ since $\boldsymbol{\Theta}$ is symmetric). Equation (15.69) proves the Lemma, as looking in the diagonal of the matrices of (15.69), one gets (because $\mathbf{D}$ is diagonal):

$$\boldsymbol{diag}(\boldsymbol{\Lambda}) = -a\boldsymbol{diag}(\mathbf{T}^\top\mathbf{A}\mathbf{T}), \quad (15.70)$$

which gives us the variation in eigenvalues (15.65), while looking outside the diagonal in (15.69), one immediately gets matrix **B** (15.66) as indeed (15.69) becomes in this case for row i, column j:

$$0 = d_{ii} b_{ij} - d_{jj} b_{ij} - a \boldsymbol{t}_i^\top \mathbf{A} \boldsymbol{t}_j . \tag{15.71}$$

When $d_{ii} \neq d_{jj}$, this leads to (15.67), as claimed.

Case 2 some eigenvalues have geometric multiplicity greater than one. Assume now without loss of generality that $\mathfrak{g}(d_{kk}) = 2$, with $d_{kk} = d_{ll}$, for some $1 \leq k \neq l \leq d$. (15.71) shows that $\boldsymbol{t}_k^\top \mathbf{A} \boldsymbol{t}_l = \boldsymbol{t}_l^\top \mathbf{A} \boldsymbol{t}_k = 0$, which implies that **A** projects vectors into the space spanned by eigenvectors $\{\boldsymbol{t}_i\}_{i \neq k,l}$, so that $\{\boldsymbol{t}_k, \boldsymbol{t}_l\}$ generates the null space of **A**. Picking $i = k, l$ or $j = k, l$ in (15.71) implies $\forall i, j \neq k, l : b_{kj} = b_{lj} = b_{ik} = b_{il} = 0$. Hence, in columns k or l, **B** may only have non-zero values in rows k or l. But looking at (15.70) shows that $\lambda_{kk} = \lambda_{ll} = 0$, implying $d'_{kk} = d_{kk} = d_{ll} = d'_{ll}$. It is immediate to check from (15.63) that $\boldsymbol{t}_k$ and $\boldsymbol{t}_l$ are also eigenvectors of $\boldsymbol{\Theta} - a\mathbf{A}$. To finish-up, looking at (15.68) brings that if the remaining unknowns in columns k or l in **B** are non-zero, then $\boldsymbol{t}_k$ and $\boldsymbol{t}_l$ are collinear, which is impossible. □

Armed with this Lemma, we can prove the following Theorem, in which we use the decomposition $\mathbf{A} = \sum_{i=1}^{d} a_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$, where a_i denotes an eigenvalue with eigenvector $\boldsymbol{a}_i$.

Theorem 4. *Define $\mathfrak{e}(\boldsymbol{\Theta}) > 0$ as the minimum difference between distinct eigenvalues of $\boldsymbol{\Theta}$, and $d^\star$ the number of distinct eigenvalues of $\boldsymbol{\Theta}$. Then, under the first-order shift setting, the following holds on ς (15.61):*

$$\varsigma \leq \left(\frac{a d^{\star 2} \mathrm{Tr}\,(\mathbf{A})^3}{\mathfrak{e}(\boldsymbol{\Theta})} \right)^4 . \tag{15.72}$$

Proof sketch: We denote $\boldsymbol{v}_i$ the eigenvector in column i of **V** in (15.63). The general term of $\mathbf{V}^\top \mathbf{T}$ in row i, column j is: $\boldsymbol{v}_i^\top \boldsymbol{t}_j$, but it comes from the definition of **B** in (15.68) that $\boldsymbol{v}_i = \boldsymbol{t}_i + \sum_k b_{ki} \boldsymbol{t}_k$, which yields $\boldsymbol{v}_i^\top \boldsymbol{t}_j = b_{ji}^2$ if $i \neq j$ (and 1 otherwise); so:

$$\begin{aligned}
\varsigma &= \left\| \mathbf{I} - (\mathbf{V}^\top \mathbf{T}) \cdot (\mathbf{V}^\top \mathbf{T}) \right\|_F \\
&= \| \mathbf{B} \cdot \mathbf{B} \|_F \\
&= \sum_{\pi(i,j)} \left(\frac{a \boldsymbol{t}_i^\top \mathbf{A} \boldsymbol{t}_j}{d_{ii} - d_{jj}} \right)^4 ,
\end{aligned}$$

where $\pi(i, j)$ is the Boolean predicate $(\mathfrak{g}(d_{ii}) = 1) \wedge (\mathfrak{g}(d_{jj}) = 1) \wedge (i \neq j)$. We finally get:

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

$$\varsigma \leq \left(\sum_{\pi(i,j)} \frac{a}{\mathfrak{e}(\boldsymbol{\Theta})} \boldsymbol{t}_i^\top \mathbf{A} \boldsymbol{t}_j\right)^4$$

$$\leq \left(\sum_{\pi(i,j)} \frac{a}{\mathfrak{e}(\boldsymbol{\Theta})} \sum_{k=1}^{d} a_k |\boldsymbol{t}_i^\top \boldsymbol{a}_k||\boldsymbol{a}_k^\top \boldsymbol{t}_j|\right)^4$$

$$\leq \left(\sum_{\pi(i,j)} \frac{a}{\mathfrak{e}(\boldsymbol{\Theta})} \sum_{k=1}^{d} a_k ||\boldsymbol{a}_k||_q ||\boldsymbol{a}_k||_r\right)^4 ,$$

by virtue of Hölder inequality ($q, r \leq \infty$), using the fact that $\mathbf{T}$ is orthonormal. Taking $q = r = 2$ and simplifying yields the statement of the Theorem.

Notice that (15.72) depends only on the eigenvalues of $\boldsymbol{\Theta}$ and $\mathbf{A}$. It says that as the "gap" in the eigenvalues of the market natural allocation increases compared to the eigenvalues of the investor's allocation, the magnitude of the interaction term decreases. Thus, the risk premium tends to depend mainly on the discrepancies (market vs investor) between "spectral" allocations for each asset, which is the separable term in (15.52).

15.7 Conclusion

In this paper, we have first proposed a generalization of Markowitz' mean-variance model, in the case where returns are not supposed anymore to be Gaussian, but are rather distributed according to exponential families of distributions with matrix arguments. Information geometry suggests that this step should be tried [2]. Indeed, because the duality collapses in this case [2], the Gaussian assumption makes that the expectation and natural parameter spaces are *identical*, which, in financial terms, represents the identity between the space of returns and the space of allocations. This, in general, can work at best only when returns are non-negative (unless short sales are allowed). Experiments suggest that the generalized model may be more accurate to spot peaks of premia, and alert investors on important market events. Our model generalizes one that we recently published, which basically uses plain Bregman divergences on vectors, which we used to learn portfolio based on their certainty equivalent [20]. The matrix extension of the model reveals interesting and non trivial roles for the two parts of the diagonalization of allocations matrices in the risk premium: the premium can indeed be split into a *separable* part which computes a premium over the spectral allocation, thus being a plain (vector) Bregman divergence part like in our former model ([20]), *plus* a non separable part which computes an interaction between stocks due to the transition matrices. We have also proposed in this paper an analysis of the magnitude of this interaction term.

299664_1_En_15_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/5/2012 Pages: 402 Layout: T1-Standard

Our model relies on Bregman matrix divergences that we have compared with others that have been previously defined elsewhere. In the general case, not restricted to allocation (SPD) matrices, our definition presents the interest to split the divergence between a separable divergence, and terms that can be non-zero when the argument matrices are not symmetric, or do not share the same transition matrices.

We have also defined Bregman matrix divergences that rely on functional composition of generators, and obtained a generalization of Bregman matrix divergences for q-norms used elsewhere [13]. We have shown that properties of the usual q-norm Bregman divergences can be generalized to our so-called Bregman–Schatten divergences. We have also proposed an on-line learning algorithm to track efficient portfolios in our matrix mean-divergence model with Bregman–Schatten divergences. The algorithm has been devised and analyzed in the setting of symmetric positive definite matrices for allocations. The algorithm generalizes conventional vector-based q-norm algorithms. Theoretical bounds for risk premia exhibit penalties that have the same flavor as those already known in the framework of supervised learning [15]. Like most of the bounds in the supervised learning literature, they are not directly applicable: in particular, we have to know ν_* beforehand for Theorem 2 to be applicable, or at least a lowerbound $\nu_\circ$ (hence, we would typically fix $\nu_\circ^{-1} \gg 1$).

From a learning standpoint, rather than finding prescient and non adaptive strategies like in constant rebalanced portfolio selection [10], on-line learning in the mean-divergence model rather aims at finding non prescient and adaptive strategies yielding efficient portfolios. This, we think, may constitute an original starting point for further works on efficient portfolio selection, with new challenging problems to solve, chief among them learning about investor's risk aversion parameters.

Acknowledgments The authors wish to thank the reviewers for useful comments, and gratefully acknowledge the support of grant ANR-07-BLAN-0328-01.

References

1. Amari, S.I.: Natural gradient works efficiently in learning. Neural Comput. **10**, 251–276 (1998)
2. Amari, S.I., Nagaoka, H.: Methods of Information Geometry. Oxford University Press, Oxford (2000)
3. Banerjee, A., Guo, X., Wang, H.: On the optimality of conditional expectation as a bregman predictor. IEEE Trans. Inf. Theory **51**, 2664–2669 (2005)
4. Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**, 1705–1749 (2005)
5. Borodin, A., El-Yaniv, R., Gogan, V.: Can we learn to beat the best stock. In: NIPS*16, pp. 345–352. (2003)
6. Bourguinat, H., Briys, E.: L'Arrogance de la Finance: comment la Théorie Financière a produit le Krach (The Arrogance of Finance: how Financial Theory made the Crisis Worse). La Découverte (2009)
7. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comp. Math. Math. Phys. **7**, 200–217 (1967)

8. Briys, E., Eeckhoudt, L.: Relative risk aversion in comparative statics: comment. Am. Econ. Rev. **75**, 281–283 (1985)
9. Chavas, J.P.: Risk Analysis in Theory and Practice. (Academic Press Advanced Finance) Academic press, London (2004)
10. Cover, T.M.: Universal portfolios. Math. Finance **1**, 1–29 (1991)
11. Dhillon, I., Sra, S.: Generalized non-negative matrix approximations with Bregman divergences. In: NIPS*18 (2005)
12. Dhillon, I., Tropp, J.A.: Matrix nearness problems with Bregman divergences. SIAM J. Matrix Anal. Appl. **29**, 1120–1146 (2007)
13. Duchi, J.C., Shalev-Shwartz, S., Singer, Y., Tewari, A.: Composite objective mirror descent. In: Proceedings of the 23^{rd} COLT, pp. 14–26. (2010)
14. Even-Dar, E., Kearns, M., Wortman, J.: Risk-sensitive online learning. In: 17^{th} ALT, pp. 199–213. (2006)
15. Kivinen, J., Warmuth, M., Hassibi, B.: The p-norm generalization of the LMS algorithm for adaptive filtering. IEEE Trans. SP **54**, 1782–1793 (2006)
16. Kulis, B., Sustik, M.A., Dhillon, I.S.: Low-rank kernel learning with Bregman matrix divergences. J. Mach. Learn. Res. **10**, 341–376 (2009)
17. Markowitz, H.: Portfolio selection. J. Finance **6**, 77–91 (1952)
18. von Neumann, J., Morgenstern, O.: Theory of games and economic behavior. Princeton University Press, Princeton (1944)
19. Nock, R., Luosto, P., Kivinen, J.: Mixed Bregman clustering with approximation guarantees. In: 23^{rd} ECML, pp. 154–169. Springer, Berlin (2008)
20. Nock, R., Magdalou, B., Briys, E., Nielsen, F.: On Tracking Portfolios with Certainty Equivalents on a Generalization of Markowitz Model: the Fool, the Wise and the Adaptive. In: Proceedings of the 28th International Conference on Machine Learning, pp. 73–80. Omnipress, Madison (2011)
21. Ohya, M., Petz, D.: Quantum Entropy and Its Use. Springer, Heidelberg (1993)
22. Petz, D.: Bregman divergence as relative operator entropy. Acta Math. Hungarica **116**, 127–131 (2007)
23. Pratt, J.: Risk aversion in the small and in the large. Econometrica **32**, 122–136 (1964)
24. Trefethen, L.N.: Numerical Linear Algebra. SIAM, Philadelphia (1997)
25. Tsuda, K., Rätsch, G., Warmuth, M.: Matrix exponentiated gradient updates for on-line learning and Bregman projection. J. Mach. Learn. Res. **6**, 995–1018 (2005)
26. Warmuth, M., Kuzmin, D.: Online variance minimization. In: 19th COLT, pp. 514–528. (2006)