Entropy Approach to Incorporate Fat Tailed Constraints in Financial Models

Santanu Dey Sandeep Juneja

Tata Institute of Fundamental Research
Mumbai, India

dsantanu@tcs.tifr.res.in juneja@tifr.res.in

May 27, 2010

Abstract

In the existing financial literature, entropy based ideas have been proposed in portfolio optimization and in model calibration for options pricing. The abstracted problem corresponds to finding a probability measure that minimizes Kullbach- Leibler (KL) distance with respect to a known measure while it satisfies certain moment constraints on functions of underlying assets. In this paper, we show that under KL distance, the optimal solution may not exist when constraints involve fat tailed distributions ubiquitous in financial practice. We note that this drawback may be corrected if 'polynomial-divergence' entropy distance is used. We discuss existence and uniqueness issues related to this new optimization problem as well as the nature of the optimal solution under different objectives. We also identify the optimal solution structure under KL distance as well as polynomial divergence when the associated constraints include those on marginal distribution of functions of underlying assets. These results are applied to a simple problem of model calibration to options prices as well as to portfolio modeling in Markowitz framework, where we note that a reasonable view that a particular portfolio of assets has heavy tailed losses may lead to fatter and more reasonable tail distributions of all assets.

1 Introduction

Entropy based ideas have found two popular applications in finance over the last fifteen years. The first involves portfolio optimization where these are used (see Meucci [23]) to arrive at a 'posterior' probability measure that is closest to the specified 'prior' probability measure and satisfies expert views modeled as constraints on certain moments associated with the posterior probability measure. The second involves calibrating the risk neutral probability measure used for pricing options (see, e.g., Buchen and Kelly in [5], Avellaneda et al. [3]). Here entropy based ideas are used to arrive at a probability measure that correctly prices given liquid options while again being closest to a specified 'prior' probability measure.

In these works, Kullbach-Leibler (KL) distance is used as a measure of distance between probability measures. One advantage of KL distance is that under mild conditions, the posterior probability measure exists and has an elegant representation when the underlying model corresponds to a distribution of light tailed random variables. However, we note this is no longer true when the underlying random variables may be fat-tailed, as is often the case in finance and insurance settings. One of our key contributions is to note that when probability distance measures corresponding to 'polynomial divergence' (defined later) are used in place of KL distance, under technical conditions, the posterior probability measure exists and has an elegant representation even when the underlying random variables may be fat-tailed. Thus, this provides a reasonable way to incorporate restrictions in the presence of fat tails. Our another contribution is that we discuss how to arrive at a posterior probability measure when the constraints on this measure are of a general nature that include involving specification of marginal distributions of functions of underlying random variables. For instance, in portfolio optimization settings, an expert may have a view that certain index of stocks has a fat-tailed t-distribution and is looking for a posterior distribution that satisfies this requirement while being closest to a prior model that may, for instance, be based on historical data.

The evolving literature on updating models for portfolio optimization builds upon the pioneering work of Black and Litterman [4] (BL). BL consider variants of Markowitz's model where the subjective views of portfolio managers are used as constraints to update models of the market using ideas from Bayesian analysis. Their work focused on Gaussian framework with views restricted to linear combinations on expectations of returns from different securities. Since then a number of variations and improvements have been suggested (see, e.g., [25], [26], [27], [21], [22], [24]). Recently, Meucci [23] proposed entropy pooling (EP) where the original model can involve general distributions and views can be on any characteristic of the probability model. Specifically, he focuses on approximating the original distribution of asset returns by a discrete one generated via Monte-Carlo sampling (or from data). Then a convex programming problem is solved that adjusts weights to these samples so that they minimize the KL distance from the original sampled distribution while satisfying the view constraints. These samples with updated weights are then used for portfolio optimization.

Buchen and Kelly in [5] used the entropy approach to calibrate one-period asset pricing models by selecting a pricing measure that correctly prices a set of benchmark instruments while minimizing KL distance from a prior specified model, that may, for instance be estimated from historical data. Avellaneda et al. [3] applied the entropy based ideas to calibrate volatility surfaces of multi period option pricing model. Avellaneda [2] calibrates continuous time asset pricing models using similar ideas. Glasserman and Yu [14] apply similar ideas to calibrate models to market data by adjusting weights of Monte-Carlo generated samples.

As mentioned earlier, in this paper we build upon ideas proposed by Meucci [23] and Buchen and Kelly [5]. We first note the well known result that for views expressed as finite number of moment constraints, the optimal solution to the KL distance minimization can be characterized as a probability measure obtained by suitably exponentially twisting the original measure. This measure is known in literature as the Gibbs measure and our analysis is based on the well known ideas involved in Gibbs conditioning principle (see, for instance, [9]). We further note that such a characterization may fail when the underlying distributions are fat-tailed in the sense that they do not have appropriate exponential moments. We show that one reasonable way to get a good change of measure that incorporates views[1] in this setting is by replacing KL distance by a suitable 'polynomial-divergence' as an objective in our optimization problem. We characterize the unique optimal solution in this setting. Our definition of polynomial-divergence is a special case of a more general concept of *f-divergence* introduced by Csiszar in [7]. Importantly, polynomial-divergence is monotonically increasing function of the well known *Tsallis Entropy* [30] and *Renyi Entropy* [28], and moreover, under appropriate limit converges to KL distance.

As also mentioned earlier, we consider the case where the expert views may correspond to probability distribution of functions of random variables involved. In our analysis, we show that such views, in addition to views on moments of functions of underlying random variables are easily incorporated. In particular, under technical conditions, we characterize the optimal solution with these general constraints, when the objective may be KL distance or polynomial-divergence and show the uniqueness of the resulting optimal probability measure in each case.

As an illustration, we apply these results to portfolio modeling in Markowitz framework where the returns from finite number of assets have a multi-variate Gaussian distribution and expert view is that a certain portfolio of returns is fat-tailed. We show that in the resulting probability measure, under mild conditions, all assets are similarly fat-tailed. Thus, this becomes a reasonable way to incorporate realistic tail behavior in a portfolio of assets. Generally speaking, the proposed approach may be useful in better risk management by building conservative tail views in mathematical models.

Note that a key reason to propose polynomial divergence is that it provides a tractable and elegant way to arrive at a reasonable updated distribution close to the given prior distribution while incorporating constraints and views even when fat-tailed distributions are involved. It is natural to try and understand the influence of the choice of objective function on the resultant optimal probability measure. We address this issue for a simple example where we compare the three reasonable objectives: The total variational distance, KL distance or polynomial-divergence and discuss the differences in the resulting solutions. To shed further light on this, we also observe that when views are expressed as constraints on probability values of disjoint sets, the optimal solution is the same in all three cases. Furthermore, it has a simple representation.

A brief historical perspective on related entropy based literature may be in order: This concept was first introduced by Gibbs in the framework of classical theory of thermodynamics (see [13]) where entropy was defined as a measure of *disorder* in thermodynamical systems. Later, Shannon [29] (see also [20]) proposed that entropy could be interpreted as a measure of *missing information* of a random system. Jayenes [16], [17] further developed this idea in the framework of statistical inferences and gave the mathematical formulation of *principle of maximum entropy* (PME), which states that given partial information/views or constraints about an unknown probability distribution, among all the distributions that satisfy these restrictions, the distribution that maximizes the entropy is the one that is least prejudiced, in the sense of being minimally committal to missing information. When a prior probability distribution, say, μ is given, one can extend above principle to *principle of minimum cross entropy* (PMXE),

[1] In this article, we often use 'views' or 'constraints' interchangeably

which states that among all probability measures which satisfy a given set of constraints, the one with minimum relative entropy (or the KL distance) with respect to μ, is the one that is maximally committal to the prior μ. See [19] for numerous applications of PME and PMXE in diverse fields of science and engineering. See [8],[1],[10],[15] and [18] for axiomatic justifications for Renyi and Tsallis entropy.

This article is organized as follows: In Section 2, we outline the mathematical framework and characterize the optimal probability measure that minimizes the KL distance with respect to the original probability measure subject to views expressed as moment constraints of specified functions. In Section 3, we show through an example that KL distance may be inappropriate objective function in the presence of fat-tailed distributions. We then define polynomial-divergence and characterize the optimal probability measure that minimizes this divergence subject to constraints on moments. The uniqueness of the optimal measure, when it exists, is proved under technical assumptions. We also discuss existence of the solution in a simple setting. In Section 4, we extend the methodology to incorporate views on marginal distributions of some random variables, along with views on moments of functions of underlying random variables. We characterize the optimal probability measures that minimize KL distance and polynomial-divergence in this setting and prove the uniqueness of the optimal measure when it exists. In Section 5, we apply our results to the portfolio problem in the Markowitz framework and develop explicit expressions for the posterior probability measure. We also show how a view that a portfolio of assets has a fat-tailed distribution renders a similar fat-tailed marginal distribution to all assets correlated to this portfolio. Section 6 is devoted to comparing qualitative differences on a simple example in the resulting optimal probability measures when the objective function is KL distance, polynomial-divergence and total variational distance. In this section, we also note that when views are on probabilities of disjoint sets, all three objectives give identical results. We numerically test our proposed algorithms on practical examples in Section 7. Finally, we end in Section 8 with a brief conclusion. All but the simplest proofs are relegated to the Appendix.

2 Incorporating Views using KL Distance

Some notation and basic concepts are needed to support our analysis. Let $(\Omega, \mathcal{F}, \mu)$ denote the underlying probability space. Let $\mathcal{P}$ be the set of all probability measures on$(\Omega, \mathcal{F})$. For any $\nu \in \mathcal{P}$ the *relative entropy* of ν w.r.t μ or equivalently the *Kullback-Leibler distance* or *I-divergence* of ν w.r.t μ is defined as

$$D(\nu \mid\mid \mu) := \int \log\left(\frac{d\nu}{d\mu}\right) d\nu$$

if ν is absolutely continuous with respect to μ and $log(\frac{d\nu}{d\mu})$ is integrable, and $D(\nu \mid\mid \mu) = +\infty$ otherwise. See, for instance [6], for concepts related to relative entropy.

Let $\mathcal{P}(\mu)$ be the set of all probability measures which are absolutely continuous w.r.t. μ. Let $\psi : \Omega \to \mathbb{R}$ be a measurable function such that $\int |\psi| e^{\psi} d\mu < \infty$. Let

$$\Lambda(\psi) := \log \int e^{\psi}\, d\mu \ \in (-\infty, +\infty]$$

denote the logarithmic moment generating function of ψ w.r.t μ.

Then, it is well known that

$$\Lambda(\psi) = \sup_{\nu \in \mathcal{P}(\mu)} \{ \int \psi\, d\nu - D(\nu \mid\mid \mu) \}.$$

Furthermore, this supremum is attained at ν^* given by:

$$\frac{d\nu^*}{d\mu} = \frac{e^{\psi}}{\int e^{\psi}\, d\mu}. \tag{1}$$

(see, for instance, [6], [19], [11]).

In our optimization problem we look for a probability measure $\nu \in \mathcal{P}(\mu)$ that minimizes the KL distance w.r.t. μ. We restrict our search to probability measures that satisfy moment constraints $\int g_i\, d\nu \geq c_i$, and/or $\int g_i\, d\nu = c_i$, where each g_i is a measurable function. For instance, views on probability of certain sets can be modeled by setting g_i's as indicator functions of those sets. If our underlying space supports random variables $(X_1, \ldots, X_n)$ under the probability measure μ, one may set $g_i = f_i(X_1, \ldots, X_n)$ so that the associated constraint is on the expectation of these functions.

Formally, our optimization problem $\mathbf{O_1}$ is:

$$\min_{\nu \in \mathcal{P}(\mu)} \int log\left(\frac{d\nu}{d\mu}\right)\, d\nu \tag{2}$$

subject to the constraints:

$$\int g_i\, d\nu \geq c_i, \tag{3}$$

for $i = 1, \ldots, k_1$ and

$$\int g_i\, d\nu = c_i, \tag{4}$$

for $i = k_1 + 1, \ldots, k$. Here k_1 can take any value between 0 and k.

The solution to this is characterized by the following assumption:

Assumption 1 *There exist $\lambda_i \geq 0$ for $i = 1, \ldots, k_1$, and $\lambda_{k_1+1}, ..., \lambda_k \in \mathbb{R}$ such that*

$$\int e^{\sum_i \lambda_i g_i}\, d\mu < \infty$$

and the probability measure ν^0 given by

$$\nu^0(A) = \int_A \frac{e^{\sum_i \lambda_i g_i}\, d\mu}{\int e^{\sum_i \lambda_i g_i}\, d\mu} \tag{5}$$

for all $A \in \mathcal{F}$ satisfies the constraints (3) and (4). Furthermore, the complementary slackness conditions

$$\lambda_i(c_i - \int g_i\, d\nu) = 0,$$

hold for $i = 1, \ldots, k_1$.

The following theorem follows:

Theorem 1 *Under Assumption(1), ν^0 is an optimal solution to $\mathbf{O_1}$.*

This theorem is well known and a proof using Lagrange multiplier method can be found in [6], [19], [11], [5], [2]. For completeness we sketch the proof below.

Proof of Theorem 1: $\mathbf{O_1}$ is equivalent to maximizing $-D(\nu \mid\mid \mu) = -\int log(\frac{d\nu}{d\mu})\, d\nu$ subject to the constraints (3) and (4). The Lagrangian for the above maximization problem is:

$$\begin{aligned}\mathcal{L} &= \sum_i \lambda_i \left(\int g_i \, d\nu - c_i \right) + (-D(\nu \mid\mid \mu)) \\ &= \sum_i \int \psi \, d\nu - D(\nu \mid\mid \mu) - \sum_i \lambda_i c_i,\end{aligned}$$

where $\psi = \sum_i \lambda_i g_i$. Then by (1) and the preceding discussion, it follows that ν^0 maximizes $\mathcal{L}$. By Lagrangian duality, due to Assumption 1, ν^0 also solves $\mathbf{O_1}$. □

Note that to obtain the optimal distribution by formula (5), we must solve the constraint equations for the Lagrange multipliers $\lambda_1, \lambda_2, \ldots, \lambda_k$. The constraint equations with its explicit dependence on λ_i's can be written as:

$$\frac{\int g_j e^{\sum_i \lambda_i g_i} \, d\mu}{\int e^{\sum_i \lambda_i g_i} \, d\mu} = c_j \text{ for } j = 1, 2, \ldots, k. \tag{6}$$

This is a set of k nonlinear equations in k unknowns $\lambda_1, \lambda_2, \ldots, \lambda_k$, and typically would require numerical procedures for solution. In general, little can be said about existence of a solution. It is easy to see that if the constraint equations are not consistent then no solution exists. On the other hand, when a solution does exist, it is helpful for applying numerical procedures, to know if it is unique. It can be shown that the Jacobian matrix of the set of equation (6) is given by the variance-covariance matrix of $g_1, g_2, \ldots, g_k$ under the measure given by (5). It follows that if no non-zero linear combination of $g_1, g_2, \ldots, g_k$ has zero variance under the measure given by (5), then solution to (6) is unique. Details can be found in [5]. In Section 4.2, we prove this in a more general setting where the marginal distributions of random variables may be constrained. We end this section with a small example.

Example 1 Suppose that under μ, random variables $X = (X_1, \ldots, X_n)$ have a multivariate normal distribution $N(a, \Sigma)$, that is, with mean $a \in \mathbb{R}^n$ and variance covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. If constraints correspond to their mean vector equaling $\hat{a}$, then this can achieved by a new probability measure ν^0 obtained by exponentially twisting μ by a vector $\lambda \in \mathbb{R}^n$ such that

$$\lambda = (\Sigma^{-1})^T (\hat{a} - a).$$

Then, under ν^0, X is $N(\tilde{a}, \Sigma)$ distributed.

3 Incorporating Views using Polynomial-Divergence

In this section, we first note through a simple example involving fat-tailed distribution that optimal solution under KL distance may not exist in certain settings. In fact, in this simple setting, one can obtain a solution that is arbitrarily close to the original distribution in the sense of KL distance. However, the form of such solutions may be inelegant and not reasonable as a model in financial settings. This motivates the use of other notions of probability distance measures as objectives such as f-divergence (introduced by Csiszar [7]). We first define general

f-divergence and later concentrate on the case where f has the form $f(u) = u^{\beta+1}$, $\beta > 0$. We refer to this as polynomial-divergence and note its relation with Tsallis entropy, Renyi entropy, and Kullback-Leibler distance. We then characterize the optimal solution under polynomial-divergence.

3.1 Polynomial Divergence

Example 2 Suppose that under μ, non-negative random variable X has a Pareto distribution with probability density function

$$f(x) = \frac{\alpha - 1}{(1+x)^{\alpha}}\ ,\ x \geq 0,\ \ \alpha > 2.$$

The mean under this pdf equals $1/(\alpha - 2)$. Suppose the view is that the mean should equal $c > 1/(\alpha - 2)$. It is well known and easily checked that

$$\frac{\int xe^{\lambda x} f(x)dx}{\int e^{\lambda x} f(x)dx}$$

is an increasing function of λ that equals ∞ for $\lambda > 0$. Hence, Assumption 1, does not hold for this example. Similar difficulty arises with other fat-tailed distributions such as Log-normal and t-distribution.

To shed further light on Example 2, for $M > 0$, consider a probability distribution

$$f_\lambda(x) = \frac{\exp(\lambda x) f(x) I_{[0,M]}(x)}{\int_0^M \exp(\lambda x) f(x)dx}.$$

Where $I_A(\cdot)$ denote the indicator function of the set A, that is, $I_A(x) = 1$ if $x \in A$ and 0 otherwise. Let λ_M denote the solution to

$$\int_0^\infty x f_\lambda(x)dx = c\ (> 1/(\alpha - 2)).$$

Proposition 1 *The sequence $\{\lambda_M\}$ and the the KL distance $\int \log\left(\frac{f_{\lambda_M}(x)}{f(x)}\right) \tilde{f}_M(x)dx$ both converge to zero as $M \to \infty$.*

Solutions such as f_{λ_M} above are typically not representative of many applications, motivating the need to have alternate methods to arrive at reasonable posterior measures that are close to μ and satisfy constraints such as (3) and (4) while not requiring that the optimal solution be obtained using exponential twisting.

We now address this issue using polynomial-divergence. Consider below f-divergence as introduced by Csiszar in [7]:

Definition 1 *Let $f : (0, \infty) \to \mathbb{R}$ be a strictly convex function. The f-divergence of a probability measure ν w.r.t. another probability measure μ equals*

$$I_f(\nu \mid\mid \mu) := \int f\left(\frac{d\nu}{d\mu}\right) d\mu$$

if ν is absolutely continuous and $f(\frac{d\nu}{d\mu})$ is integrable w.r.t. μ. Otherwise we set $I_f(\nu \mid\mid \mu) = \infty$.

Note that KL distance corresponds to the case $f(u) = u \log u$. Other popular examples of f include

$$f(u) = -\log u, \;\; f(u) = u^{\beta+1}, \;\; \beta > 0, \;\; f(u) = e^u.$$

In this section we consider $f(u) = u^{\beta+1}$, $\beta > 0$ and refer to it as *polynomial-divergence.* That is, we focus on

$$I_\beta(\nu \,||\, \mu) := \int \left(\frac{d\nu}{d\mu}\right)^{\beta} d\nu = \int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu.$$

It is easy to see using Jensen's inequality that

$$\min_{\nu \in \mathcal{P}(\mu)} \int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu$$

is achieved by $\nu = \mu$.

Our optimization problem $\mathbf{O_2}(\beta)$ may be stated as:

$$\min_{\nu \in \mathcal{P}(\mu)} \int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu \tag{7}$$

subject to (3) and (4).

Remark 1 *Relation with Tsallis Entropy and Renyi Entropy:* Let α and γ be a positive real numbers. The relative Tsallis entropy with index α of ν w.r.t. μ equals

$$S_\alpha(\nu \,||\, \mu) := \int \frac{\left(\frac{d\nu}{d\mu}\right)^{\alpha} - 1}{\alpha} d\nu$$

if ν is absolutely continuous w.r.t. μ and the integral is finite. Otherwise, $S_\alpha(\nu \,||\, \mu) = \infty$. See, e.g., [30].

The relative Renyi entropy of order γ of ν w.r.t. μ equals

$$H_\gamma(\nu \,||\, \mu) := \frac{1}{\gamma - 1} \log\left(\int \left(\frac{d\nu}{d\mu}\right)^{\gamma-1} d\nu\right) \quad \text{when } \gamma \neq 1 \text{ and}$$

$$H_1(\nu \,||\, \mu) := \int \log\left(\frac{d\nu}{d\mu}\right) d\nu$$

if ν is absolutely continuous w.r.t μ and the respective integrals are finite. Otherwise, $H_\gamma(\nu \,||\, \mu) = \infty$ (see, e.g, [28]).

It can be shown that as $\gamma \to 1$, $H_\gamma(\nu \,||\, \mu) \to H_1(\nu \,||\, \mu) = D(\nu \,||\, \mu)$ and as $\alpha \to 0$, $S_\alpha(\nu \,||\, \mu) \to D(\nu \,||\, \mu)$. Also, following relations are immediate consequences of the above definitions.

$$I_\beta(\nu \,||\, \mu) = 1 + \beta S_\beta(\nu \,||\, \mu)$$

$$I_\beta(\nu \,||\, \mu) = e^{\beta H_{\beta+1}((\nu||\mu)}$$

$$\lim_{\beta \to 0} \frac{I_\beta(\nu \,||\, \mu) - 1}{\beta} = D(\nu \,||\, \mu)$$

Thus, polynomial-divergence is a strictly increasing function of both relative Tsallis entropy and relative Renyi entropy. Therefore minimizing polynomial-divergence is equivalent to minimizing relative Tsallis entropy or relative Renyi entropy.

In the following assumption we specify the solution form to $\mathbf{O_2}(\beta)$:

Assumption 2 *There exist $\lambda_i \geq 0$ for $i = 1, \ldots, k_1$, and $\lambda_{k_1+1}, ..., \lambda_k \in \mathbb{R}$ such that*

$$1 + \beta \sum_l \lambda_l g_l \geq 0 \;\; a.e.(\mu). \tag{8}$$

$$\int \left(1 + \beta \sum_l \lambda_l g_l\right)^{1/\beta} d\mu < \infty, \tag{9}$$

and the probability measure ν^1 given by

$$\nu^1(A) = \int_A \frac{(1 + \beta \sum_l \lambda_l g_l)^{1/\beta} \, d\mu}{\int (1 + \beta \sum_l \lambda_l g_l)^{1/\beta} \, d\mu} \tag{10}$$

for all $A \in \mathcal{F}$ satisfies the constraints (3) and (4). Furthermore, the complementary slackness conditions

$$\lambda_l (c_l - \int g_l \, d\nu) = 0,$$

hold for $l = 1, \ldots, k_1$.

Theorem 2 *Under Assumption 2, ν^1 is an optimal solution to $\mathbf{O_2}(\beta)$.*

Existence and Uniqueness of Lagrange multipliers: To obtain the optimal distribution by formula (10), we must solve the set of k nonlinear equations given by:

$$\frac{\int g_j \left(1 + \beta \sum_{l=1}^k \lambda_l g_l\right)^{\frac{1}{\beta}} d\mu}{\int \left(1 + \beta \sum_{l=1}^k \lambda_l g_l\right)^{\frac{1}{\beta}} d\mu} = c_j \;\; \text{for } j = 1, 2, \ldots, k \tag{11}$$

for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$. In view of Assumption 2, a solution $\boldsymbol{\lambda}$ is called *feasible* if it satisfies (8) and (9) and (11). A feasible solution $\boldsymbol{\lambda}$ is called *strongly feasible* to (11) if it further satisfies the following two conditions:

$$\int \left(1 + \beta \sum_{l=1}^k \lambda_l g_l\right)^{\frac{1}{\beta}+1} d\mu < \infty$$

and

$$1 + \beta \sum_{l=1}^k \lambda_l c_l > 0.$$

Theorem 3 states sufficient conditions under which a strongly feasible solution to (11) is unique.

Theorem 3 *Suppose that the variance-covariance matrix of $g_1, g_2, \ldots, g_k$ under any measure $\nu \in \mathcal{P}(\mu)$ is positive definite, or equivalently, that for any $\nu \in \mathcal{P}(\mu)$, $\sum_{l=1}^k a_l g_l = c \; a.e.\,(\nu)$, for some constants c and $a_1, a_2, \ldots, a_k$ implies $c = 0$ and $a_i = 0$ for all $i = 1, 2, \ldots, k$. Then, if a strongly feasible solution to (11) exists, it is unique.*

Proposition 2 below shows the existence of a solution to $\mathbf{O_2}(\beta)$ under a single random variable, single constraint settings. We then apply this to a few specific examples.

Note that for any random variable X, a non-negative function g, and an integer $n > 0$, $E[g(X)^{n+1}] \geq E[g(X)]E[g(X)^n]$ since random variables $g(X)$ and $g(X)^n$ are positively associated and hence have non-negative covariance.

Proposition 2 *Consider a random variable X with pdf f, and a function $g : \mathbb{R}^+ \to \mathbb{R}^+$ such that $E[g(X)^{n+1}] < \infty$ for a positive integer n. Further suppose that $E[g(X)^{n+1}] > E[g(X)]E[g(X)^n]$. Then the optimization problem:*

$$\min_{\tilde{f}\in\mathcal{P}(f)} \int_0^\infty \left(\frac{\tilde{f}(x)}{f(x)}\right)^{1+1/n} f(x)\,dx$$

$$\textit{subject to: } \frac{\tilde{E}[g(X)]}{E[g(X)]} = a$$

has a unique solution for $a \in \left(1, \frac{E[g(X)^{n+1}]}{E[g(X)]E[g(X)^n]}\right)$, given by

$$\tilde{f}(x) = \frac{\left(1+\frac{\lambda}{n}g(x)\right)^n f(x)}{\sum_{k=0}^n n^{-k}\binom{n}{k}E[g(X)^k]\lambda^k}\ , x \geq 0,$$

where λ is a positive root of the polynomial

$$\sum_{k=0}^n n^{-k}\binom{n}{k}\left\{E[g(X)^{k+1}] - aE[g(X)]E[g(X)^k]\right\}\lambda^k = 0. \tag{12}$$

As we note in the proof in the Appendix, the uniqueness of the solution follows from Theorem 3.

Example 3 Suppose X is log-normally distributed with parameters (μ, σ^2) (that is, $\log X$ has normal distribution with mean μ and variance σ^2). Then, its density function

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}}\exp\{-\frac{(\log x-\mu)^2}{2\sigma^2}\}\ , x \geq 0$$

For the constraint

$$\frac{\tilde{E}(X)}{E(X)} = a, \tag{13}$$

first consider the case $\beta = 1/n = 1$. Then, the probability distribution minimizing the polynomial-divergence is given by:

$$\tilde{f}(x) = \frac{(1+\lambda x)f(x)}{1+\lambda E(X)} \quad x \geq 0.$$

From the constraint equation we have:

$$a = \frac{\tilde{E}(X)}{E(X)} = \frac{1}{E(X)+\lambda E(X)^2}\int_0^\infty x(1+\lambda x)f(x)dx = \frac{E(X)+\lambda E(X^2)}{E(X)+\lambda E(X)^2},$$

or

$$\lambda = \frac{E(X)(a-1)}{E(X^2) - aE(X)^2} = \frac{a-1}{e^{\mu+\sigma^2/2}(e^{\sigma^2} - a)}$$

Since $\frac{E(X)+\lambda E(X^2)}{E(X)+\lambda E(X)^2}$ increases with λ and converges to $\frac{E(X^2)}{E(X)^2} = e^{\sigma^2}$ as $\lambda \to \infty$, it follows that our optimization problem has a solution if $a \in [1, e^{\sigma^2})$. Thus if $a > e^{\sigma^2}$ and $\beta = 1$, Assumption 2 cannot hold.

It is easily checked that $\frac{E(X^{n+1})}{E(X)E(X^n)} = e^{n\sigma^2}$. Then, for $\beta = \frac{1}{n}$, a solution always exists for $a \in \left[1, e^{n\sigma^2}\right)$.

Example 4 Suppose that rv X has a Gamma distribution with density function

$$f(x) = \frac{\theta^\alpha x^{\alpha-1} e^{-\theta x}}{\Gamma(\alpha)} \ , x \geq 0$$

and as before, the constraint is given by (13). Then, it is easily seen that $\frac{E(X^{n+1})}{E(X)E(X^n)} = 1 + \frac{n}{\alpha}$, so that a solution with $\beta = \frac{1}{n}$ exists for $a \in \left[1, 1 + \frac{n}{\alpha}\right)$.

Example 5 Suppose X has a Pareto distribution with probability density function:

$$f(x) = \frac{\alpha - 1}{(x+1)^\alpha} \ , x \geq 0 \, ,$$

and as before, the constraint is given by (13). Then, it is easily seen that

$$\frac{E[X^{n+1}]}{E[X]E[X^n]} = \frac{(\alpha - n - 1)(\alpha - 2)}{(\alpha - n - 2)(\alpha - 1)}.$$

As in previous examples, we see that a probability distribution minimizing the polynomial-divergence with $\beta = 1/n$ with $n < \alpha - 2$ exists when $a \in \left[1, \frac{(\alpha-n-1)(\alpha-2)}{(\alpha-n-2)(\alpha-1)}\right)$.

Hence, in the above examples, for any $a \geq 1$, Assumption 2 has a solution for all β sufficiently small.

4 Incorporating Constraints on Marginal Distributions

Next we state and prove the analogue of Theorem 1 and 2 when there is a constraint on marginal distribution of few components of the given random vector. Later in Remark 2 we discuss how this generalizes to the case where the constraints involve moments and marginals of functions of the given random vector. Let $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors having joint law μ which is given by joint probability density function $f(\mathbf{x}, \mathbf{y})$. Recall that $\mathcal{P}(\mu)$ is the set of all probability measures which are absolutely continuous w.r.t. μ. If $\nu \in \mathcal{P}(\mu)$ then ν is also specified by a probability density function say, $\tilde{f}(\cdot)$, such that $\tilde{f}(\mathbf{x}, \mathbf{y}) = 0$ whenever $f(\mathbf{x}, \mathbf{y}) = 0$ and $\frac{d\nu}{d\mu} = \frac{\tilde{f}}{f}$. In view of this we may formulate our optimization problem in terms of probability density functions instead of measures. Let $\mathcal{P}(f)$ denote the collection of density functions that are absolutely continuous with respect to the density f.

4.1 Incorporating views on Marginal using KL distance

Formally, our optimization problem $\mathbf{O_3}$ is:

$$\min_{\nu \in \mathcal{P}(\mu)} \int log\left(\frac{d\nu}{d\mu}\right) d\nu = \min_{\tilde{f} \in \mathcal{P}(f)} \int \log\left(\frac{\tilde{f}(\mathbf{x}, \mathbf{y})}{f(\mathbf{x}, \mathbf{y})}\right) \tilde{f}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y},$$

subject to:

$$\int_{\mathbf{y}} \tilde{f}(\mathbf{x}, \mathbf{y}) d\mathbf{y} = g(\mathbf{x}) \text{ for all } \mathbf{x}, \tag{14}$$

where $g(\mathbf{x})$ is a given marginal density function of $\mathbf{X}$, and

$$\int_{\mathbf{x},\mathbf{y}} h_i(\mathbf{x}, \mathbf{y}) \tilde{f}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} = c_i \tag{15}$$

for $i = 1, 2 \ldots, k$. For presentation convenience, in the remaining paper we only consider equality constraints on moments of functions (as in (15)), ignoring the inequality constraints. The latter constraints can be easily handled as in Assumptions (1) and (2) by introducing suitable non-negativity and complimentary slackness conditions.

Some notation is needed to proceed further. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ and $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_k)$. Further,

$$f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) := \frac{\exp(\sum_i \lambda_i h_i(\mathbf{x}, \mathbf{y})) f(\mathbf{y}|\mathbf{x})}{\int_{\mathbf{y}} \exp(\sum_j \lambda_i h_i(\mathbf{x}, \mathbf{y})) f(\mathbf{y}|\mathbf{x}) d\mathbf{y}} = \frac{\exp(\sum_i \lambda_i h_i(\mathbf{x}, \mathbf{y})) f(\mathbf{x}, \mathbf{y})}{\int_{\mathbf{y}} \exp(\sum_i \lambda_i h_i(\mathbf{x}, \mathbf{y})) f(\mathbf{x}, \mathbf{y}) d\mathbf{y}}.$$

Let $f_{\boldsymbol{\lambda}}(\mathbf{x}, \mathbf{y})$ denote the joint density function of $(\mathbf{X}, \mathbf{Y})$, $f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) \times g(\mathbf{x})$ for all $\mathbf{x}, \mathbf{y}$, and $E_{\boldsymbol{\lambda}}$ denote the expectation under $f_{\boldsymbol{\lambda}}$.

For a mathematical claim $S(x)$, we write $S(x)$ for almost all x *w.r.t.* $g(x)dx$ to mean that $m_g(x|\, S(x) \text{ is false}) = 0$, where m_g is the measure induced by the density g. That is, $m_g(A) = \int_A g(x)\, dx$ for all measurable subsets A.

Assumption 3 *There exists $\boldsymbol{\lambda} \in \mathbb{R}^k$ such that*

$$\int_{\mathbf{y}} \exp(\sum_i \lambda_i h_i(x, y)) f(\mathbf{x}, \mathbf{y}) d\mathbf{y} < \infty$$

for almost all $\mathbf{x}$ w.r.t $g(\mathbf{x})d\mathbf{x}$ and the probability density function $f_{\boldsymbol{\lambda}}$ satisfies the constraints given by (15). That is, for all $i = 1, 2, ..., k$, we have

$$E_{\boldsymbol{\lambda}}[h_i(\mathbf{X}, \mathbf{Y})] = c_i\,. \tag{16}$$

Theorem 4 *Under Assumption(3), $f_{\boldsymbol{\lambda}}(\cdot)$ is an optimal solution to $\mathbf{O_3}$.*

In Theorem 5, we develop conditions that ensure uniqueness of a solution to $\mathbf{O_3}$ once it exists.

Theorem 5 *Suppose that for almost all $\mathbf{x}$ w.r.t. $g(\mathbf{x})d\mathbf{x}$, conditional on $\mathbf{X} = \mathbf{x}$, no non-zero linear combination of the random variables $h_1(\mathbf{x}, \mathbf{Y}), h_2(\mathbf{x}, \mathbf{Y}), \ldots, h_k(\mathbf{x}, \mathbf{Y})$ has zero variance w.r.t. the conditional density $f(\mathbf{y}|\mathbf{x})$, or, equivalently, for almost all $\mathbf{x}$ w.r.t. $g(\mathbf{x})d\mathbf{x}$, $\sum_i a_i h_i(\mathbf{x}, \mathbf{Y}) = c$ almost surely ($f(\mathbf{y}|\mathbf{x})d\mathbf{y}$) for some constants c and $a_1, a_2, \ldots, a_k$ implies $c = 0$ and $a_i = 0$ for all $i = 1, 2, \ldots, k$. Then, if a solution to the constraint equations (16) exists, it is unique.*

Remark 2 Theorem 4, as stated, is applicable when the updated marginal distribution of a random sub-vector $\mathbf{X}$ of the given random vector $(\mathbf{X}, \mathbf{Y})$ is specified. More generally, by a routine change of variable technique, similar specification on a function of the given random vector can also be incorporated. We now illustrate this.

Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$ denote a random vector taking values in $S \subseteq \mathbb{R}^N$ and having a (prior) density function $f_{\mathbf{Z}}$. Suppose the constraints are as follows:

- $(v_1(\mathbf{Z}), v_2(\mathbf{Z}), \ldots, v_{k_1}(\mathbf{Z}))$ have a joint density function given by $g(\cdot)$.
- $\tilde{E}[v_{k_1+1}(\mathbf{Z})] = c_1$, $\tilde{E}[v_{k_1+2}(\mathbf{Z})] = c_2$, $\ldots$, $\tilde{E}[v_{k_2}(\mathbf{Z})] = c_{k_2-k_1}$,

where $0 \leq k_1 \leq k_2 \leq N$ and $v_1(\cdot)$, $v_2(\cdot)$, $\ldots$, $v_{k_2}(\cdot)$ are some functions on S.

If $k_2 < N$ we define $N - k_2$ functions $v_{k_2+1}(\cdot)$, $v_{k_2+2}(\cdot)$, $\ldots$, $v_N(\cdot)$ such that the function $v : S \to \mathbb{R}^N$ defined by $v(\mathbf{z}) = (v_1(\mathbf{z}), v_2(\mathbf{z}), \ldots, v_N(\mathbf{z}))$ has a nonsingular Jacobian a.e. That is,

$$J(\mathbf{z}) := \det\left(\left(\frac{\partial v_i}{\partial z_j}\right)\right) \neq 0 \text{ for almost all } \mathbf{z} \text{ w.r.t. } f_{\mathbf{Z}},$$

where we are assuming that the functions $v_1(\cdot)$, $v_2(\cdot)$, $\ldots$, $v_{k_2}(\cdot)$ allow such a construction.

Consider $\mathbf{X} = (X_1, \ldots, X_{k_1})$, where $X_i = v_i(\mathbf{Z})$ for $i \leq k_1$ and $\mathbf{Y} = (Y_1, \ldots, Y_{N-k_1})$, where $Y_i = v_{k_1+i}(\mathbf{Z})$ for $i \leq N - k_1$. Let $f(\cdot, \cdot)$ denote the density function of $(\mathbf{X}, \mathbf{Y})$. Then, by the change of variables formula for densities,

$$f(\mathbf{x}, \mathbf{y}) = f_{\mathbf{Z}}\left(w(\mathbf{x}, \mathbf{y})\right)\left[J\left(w(\mathbf{x}, \mathbf{y})\right)\right]^{-1},$$

where $w(\cdot)$ denotes the local inverse function of $v(\cdot)$, that is, if $v(\mathbf{z}) = (\mathbf{x}, \mathbf{y})$, then, $\mathbf{z} = w(\mathbf{x}, \mathbf{y})$.

The constraints can easily be expressed in terms of $(\mathbf{X}, \mathbf{Y})$ as

$$\mathbf{X} \text{ have joint density given by } g(\cdot)$$

and

$$\tilde{E}[Y_i] = c_i \ for \ i = 1, 2, \ldots, (k_2 - k_1). \tag{17}$$

Setting $k = k_2 - k_1$, from Theorem (4) it follows that the optimal density function of $(\mathbf{X}, \mathbf{Y})$ as:

$$f_{\boldsymbol{\lambda}}(\mathbf{x}, \mathbf{y}) = \frac{e^{\lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_k y_k} f(\mathbf{x}, \mathbf{y})}{\int_y e^{\lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_k y_k} f(\mathbf{x}, \mathbf{y})\, d\mathbf{y}} \times g(\mathbf{x}),$$

where λ_k's is chosen to satisfy (17).

Again by the change of variable formula, it follows that the optimal density of $\mathbf{Z}$ is given by:

$$\tilde{f}_{\mathbf{Z}}(\mathbf{z}) = f_{\boldsymbol{\lambda}}(v_1(\mathbf{z}), v_2(\mathbf{z}), \ldots, v_N(\mathbf{z}))J(\mathbf{z}) \ . \ \square$$

4.2 Incorporating constraints on Marginals using Polynomial-Divergence

Extending Theorem 4 to the case of polynomial-divergence is straightforward. We state the details for completeness. As in the case of KL distance, the following notation will simplify our exposition. Let

$$\begin{aligned} f_{\boldsymbol{\lambda},\beta}(\mathbf{y}|\mathbf{x}) : &= \frac{\left(1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y})\right)^{\frac{1}{\beta}} f(\mathbf{y}|\mathbf{x})}{\int_{\mathbf{y}}\left(1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y})\right)^{\frac{1}{\beta}} f(\mathbf{y}|\mathbf{x})d\mathbf{y}} \\ &= \frac{\left(1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y})\right)^{\frac{1}{\beta}} f(\mathbf{x},\mathbf{y})}{\int_{\mathbf{y}}\left(1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y})\right)^{\frac{1}{\beta}} f(\mathbf{x},\mathbf{y})d\mathbf{y}} . \end{aligned}$$

If the marginal of $\mathbf{X}$ is given by $g(\mathbf{x})$ then the joint density $f_{\boldsymbol{\lambda},\beta}(\mathbf{y}|\mathbf{x}) \times g(\mathbf{x})$ is denoted by $f_{\boldsymbol{\lambda},\beta}(\mathbf{x},\mathbf{y})$. $E_{\boldsymbol{\lambda},\beta}$ denotes the expectation under $f_{\boldsymbol{\lambda},\beta}(\cdot)$.

Consider the optimization problem $\mathbf{O_4}(\beta)$:

$$\min_{\tilde{f}\in\mathcal{P}(f)} \int \left(\frac{\tilde{f}(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}\right)^{\beta} \tilde{f}(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} ,$$

subject to (14) and (15).

Assumption 4 *There exists $\boldsymbol{\lambda} \in \mathbb{R}^k$ such that*

$$1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y}) \geq 0 , \tag{18}$$

for almost all $(\mathbf{x},\mathbf{y})$ w.r.t. $f(\mathbf{y}|\mathbf{x}) \times g(\mathbf{x})d\mathbf{y}d\mathbf{x}$ and

$$\int_{\mathbf{y}}\left(1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y})\right)^{\frac{1}{\beta}} f(\mathbf{x},\mathbf{y})d\mathbf{y} < \infty , \tag{19}$$

for almost all $\mathbf{x}$ w.r.t. $g(\mathbf{x})d\mathbf{x}$.

Further, the probability density function $f_{\boldsymbol{\lambda},\beta}(\mathbf{x},\mathbf{y})$ satisfies the constraints given by (15). That is, for all $i = 1, 2, ..., k$, we have

$$E_{\boldsymbol{\lambda},\beta}[h_i(\mathbf{X},\mathbf{Y})] = c_i . \tag{20}$$

Theorem 6 *Under Assumption(4), $f_{\boldsymbol{\lambda},\beta}(\cdot)$ is an optimal solution to $\mathbf{O_4}(\beta)$.*

Analogous to the discussion in Remark (2), by suitable change of variable, we can adapt the above theorem to the case where the constraints involve marginal distribution and/or moments of functions of a given random vector.

We conclude this section with a brief discussion on uniqueness of the solution to $\mathbf{O_4}(\beta)$. Any $\boldsymbol{\lambda} \in \mathbb{R}^k$ satisfying (18), (19) and (20) is called a *feasible solution* to $\mathbf{O_4}(\beta)$. A feasible solution $\boldsymbol{\lambda}$ is called *strongly feasible* if, in addition, for almost all $\mathbf{x}$ w.r.t. $g(\mathbf{x})d\mathbf{x}$ it satisfies the following conditions:

$$\int_{\mathbf{y}}\left(1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j h_j(\mathbf{x},\mathbf{y})\right)^{\frac{1}{\beta}+1} f(\mathbf{x},\mathbf{y})d\mathbf{y} < \infty \text{ for almost all } \mathbf{x} \text{ w.r.t. } g(\mathbf{x})d\mathbf{x}$$

and

$$1+\beta\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right)^{\beta}\sum_j \lambda_j c_j > 0 \text{ for almost all } \mathbf{x} \text{ w.r.t. } g(\mathbf{x})d\mathbf{x}.$$

The following theorem can be proved using similar arguments as those used to prove Theorem 3. We omit the details.

Theorem 7 *Suppose that for almost all $\mathbf{x}$ w.r.t. $(g(\mathbf{x})d\mathbf{x})$, conditional on $\mathbf{X}=\mathbf{x}$, no non-zero linear combination of $h_1(\mathbf{x},\mathbf{Y}), h_2(\mathbf{x},\mathbf{Y}),\ldots,h_k(\mathbf{x},\mathbf{Y})$ has zero variance under any measures ν absolutely continuous w.r.t $f(\mathbf{y}|\mathbf{x})$. Or equivalently, that for any measures ν absolutely continuous w.r.t $f(\mathbf{y}|\mathbf{x})$, $\sum_{l=1}^{k} a_l h_l(\mathbf{x},\mathbf{Y}) = c$ almost everywhere (ν), for some constants c and $a_1,a_2,\ldots,a_k$, implies $c=0$ and $a_l=0$ for all $l=1,2,\ldots,k$. Then, if a strongly feasible solution to (20) exists, it is unique.*

5 Portfolio Modeling in Markowitz Framework

In this section we apply the methodology developed in Section 4.1 to the Markowitz framework: Namely to the setting where there are N assets whose returns under the ‘prior distribution’ are multi-variate Gaussian. Here, we explicitly identify the posterior distribution that incorporates views/constraints on marginal distribution of some random variables and moment constraints on other random variables. As mentioned in the introduction, an important application of our approach is that if for a particular portfolio of assets, say an index, it is established that the return distribution is fat-tailed (specifically, the pdf is a regularly varying function), say with the density function g, then by using that as a constraint, one can arrive at an updated posterior distribution for all the underlying assets. Furthermore, we show that if an underlying asset has a non-zero correlation with this portfolio under the prior distribution, then under the posterior distribution, this asset has a tail distribution similar to that given by g.

Let $(\mathbf{X},\mathbf{Y}) = (X_1,X_2,\ldots,X_{N-k},Y_1,Y_2,\ldots,Y_k)$ have a N dimensional multi-variate Gaussian distribution with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_{\mathbf{x}},\boldsymbol{\mu}_{\mathbf{y}})$ and the variance-covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}. \end{pmatrix}$$

We now consider a posterior distribution that satisfies the view that:

$$X \text{ has probability density function } g(\mathbf{x}) \text{ and } \tilde{E}(\mathbf{Y}) = \mathbf{a}.$$

where $g(\mathbf{x})$ is a given probability density function on $\mathbb{R}^{N-k}$ with finite first moments along each component and $\mathbf{a}$ is a given vector in $\mathbb{R}^k$. As we discussed in Remark 2 (see also Example 7 in Section 7), when the view is on marginal distributions of linear combinations of underlying assets, and/or on moments of linear functions of the underlying assets, the problem can be easily transformed to the above setting by a suitable change of variables.

To find the distribution of $(\mathbf{X},\mathbf{Y})$ which incorporates the above views, we solve the minimization problem $\mathbf{O_5}$:

$$\min_{\tilde{f}\in\mathcal{P}(f)} \int_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^{N-k}\times\mathbb{R}^k} \log\left(\frac{\tilde{f}(\mathbf{x},\mathbf{y})}{f(\mathbf{x},\mathbf{y})}\right)\tilde{f}(\mathbf{x},\mathbf{y})\,d\mathbf{x}d\mathbf{y}$$

subject to the constraint:

$$\int_{\mathbf{y}\in\mathbb{R}^k} \tilde{f}(\mathbf{x},\mathbf{y})d\mathbf{y} = g(\mathbf{x}) \quad \forall \mathbf{x},$$

$$\int_{\mathbf{x}\in\mathbb{R}^{N-k}} \int_{\mathbf{y}\in\mathbb{R}^k} \mathbf{y}\tilde{f}(\mathbf{x},\mathbf{y})d\mathbf{y}d\mathbf{x} = \mathbf{a}. \tag{21}$$

Where $f(\mathbf{x},\mathbf{y})$ is the density of N-variate normal distribution denoted by $\mathcal{N}_N(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Proposition 3 *Under the assumption that $\Sigma_{\mathbf{xx}}$ is invertible, the optimal solution to $\mathbf{O_5}$ is given by*

$$\tilde{f}(\mathbf{x},\mathbf{y}) = g(\mathbf{x}) \times \tilde{f}(\mathbf{y}|\mathbf{x}) \tag{22}$$

where $\tilde{f}(\mathbf{y}|\mathbf{x})$ is the probability density function of

$$\mathcal{N}_k\left(\mathbf{a} + \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x} - E_g[\mathbf{X}])\,,\, \boldsymbol{\Sigma}_{\mathbf{yy}} - \boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}\right)$$

where $E_g(\mathbf{X})$ is the expectation of X under the density function g.

Tail behavior of the marginals of the posterior distribution: We now specialize to the case where $\mathbf{X}$ (also denoted by X) is a single random variable so that $N = k+1$, and Assumption 5 below is satisfied by pdf g. Specifically, $(X,\mathbf{Y})$ is distributed as $\mathcal{N}_{k+1}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ where now

$$\boldsymbol{\mu}^T = (\mu_x, \boldsymbol{\mu}_{\mathbf{y}}^T) \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \boldsymbol{\sigma}_{x\mathbf{y}}^T \\ \boldsymbol{\sigma}_{x\mathbf{y}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix}$$

where $\boldsymbol{\sigma}_{x\mathbf{y}} = (\sigma_{xy_1}, \sigma_{xy_2}, ..., \sigma_{xy_k})^T$ with $\sigma_{xy_i} = Cov(X, Y_i)$.

Assumption 5 *The pdf $g(\cdot)$ is regularly varying, that is, there exists a constant $\alpha > 1$ ($\alpha > 1$ is needed for g to be integrable) such that*

$$\lim_{t\to\infty} \frac{g(\eta t)}{g(t)} = \frac{1}{\eta^\alpha}$$

for all $\eta > 0$ (see, for instance, [12]). In addition, for any $a \in \mathbb{R}$ and $b \in \mathbb{R}^+$

$$\frac{g(b(t-s-a))}{g(t)} \le h(s) \tag{23}$$

for some non-negative function $h(\cdot)$ independent of t (but possibly depending on a and b) with the property that $Eh(Z) < \infty$ whenever Z has a Gaussian distribution.

Remark 3 Assumption 5 holds, for instance, when g corresponds to t-distribution with n degrees of freedom, that is,

$$g(s) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + \frac{s^2}{n})^{-(\frac{n+1}{2})},$$

Clearly, g is regularly varying with $\alpha = n+1$. To see (23), note that

$$\frac{g(b(t-s-a))}{g(t)} = \frac{(1+t^2/n)^{(n+1)/2}}{(1+b^2(t-s-a)^2/n)^{(n+1)/2}} \ .$$

Putting $t' = \frac{bt}{\sqrt{n}}$, $s' = \frac{b(s+a)}{\sqrt{n}}$ and $c = \frac{1}{b}$ we have

$$\frac{(1+t^2/n)}{(1+b^2(t-s-a)^2/n)} = \frac{1+c^2t'^2}{1+(t'-s')^2} \ .$$

Now (23) readily follows from the fact that

$$\frac{1+c^2t'^2}{1+(t'-s')^2} \leq max\{1, c^2\} + c^2s'^2 + c^2|s'| \text{ , for any two real numbers } s' \text{ and } t'.$$

To verify the last inequality, note that if $t' \leq s'$ then $\frac{1+c^2t'^2}{1+(t'-s')^2} \leq 1 + c^2s'^2$ and if $t' > s'$ then

$$\frac{1+c^2t'^2}{1+(t'-s')^2} = \frac{1+c^2(t'-s'+s')^2}{1+(t'-s')^2} = \frac{1+c^2(t'-s')^2}{1+(t'-s')^2} + c^2s'^2 + c^2s'\frac{2(t'-s')}{1+(t'-s')^2}$$

$$\leq max\{1, c^2\} + c^2s'^2 + c^2|s'| \, .$$

Note that if $h(x) = x^m$ or $h(x) = \exp(\lambda x)$ for any m or λ then the last condition in Assumption 5 holds.

From Proposition (3), we note that the posterior distribution of $(X, \mathbf{Y})$ is

$$\tilde{f}(x, \mathbf{y}) = g(x) \times \tilde{f}(\mathbf{y}|x)$$

where $\tilde{f}(\mathbf{y}|x)$ is the probability density function of

$$\mathcal{N}_k\left(\mathbf{a} + \left(\frac{x - E_g(X)}{\sigma_{xx}}\right)\boldsymbol{\sigma}_{x\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{yy}} - \frac{1}{\sigma_{xx}}\boldsymbol{\sigma}_{x\mathbf{y}}\boldsymbol{\sigma}_{x\mathbf{y}}^t\right)$$

where $E_g(X)$ is the expectation of X under the density function g. Let $\tilde{f}_{Y_1}$ denote the marginal density of Y_1 under the above posterior distribution. Theorem 8 states a key result of this section.

Theorem 8 *Under Assumption 5, if* $\sigma_{xy_1} \neq 0$*, then*

$$\lim_{s\to\infty} \frac{\tilde{f}_{Y_1}(s)}{g(s)} = \left(\frac{\sigma_{xy_1}}{\sigma_{xx}}\right)^{\alpha-1} . \qquad (24)$$

Note that (24) implies that

$$\lim_{x\to\infty} \frac{\int_x \tilde{f}_{Y_1}(s)ds}{\int_x g(s)ds} = \left(\frac{\sigma_{xy_1}}{\sigma_{xx}}\right)^{\alpha-1} .$$

6 Comparing Different Objectives

Given that in many examples one can use KL distance as well as polynomial-divergence as an objective function for arriving at an updated probability measure, it is natural to compare the optimal solutions in these cases. Note that the total variation distance between two probability measures μ and ν defined on $(\Omega, \mathcal{F})$ equals

$$\sup\{\mu(A) - \nu(A)|A \in \mathcal{F}\}.$$

This may also serve as an objective function in our search for a reasonable probability measure that incorporates expert views and is close to the original probability measure. This has an added advantage of being a metric (e.g., it satisfies the triangular inequality).

We now compare these three different types of objectives to get a qualitative flavor of the differences in the optimal solutions in two simple settings (a rigorous analysis in general settings may be a subject for future research). The first corresponds to the case of single random variable whose prior distribution is exponential. In the second setting, the views correspond to probability assignments to mutually exclusive and exhaustive set of events.

6.1 Exponential Prior Distribution

Suppose that the random variable X is exponentially distributed with rate α under μ. Then its pdf equals

$$f(x) = \alpha e^{-\alpha x},\ x \geq 0.$$

Now suppose that our view is that under the updated measure ν with density function $\tilde{f}$, $\tilde{E}(X) = \int x\tilde{f}(x)\, dx = 1/\gamma > 1/\alpha$.

KL Distance: When the objective function is to minimize KL distance, the optimal solution is obtained as an exponentially twisted distribution that satisfies the desired constraint. It is easy to see that exponentially twisting an exponential distribution with rate α by an amount θ leads to another exponential distribution with rate $\alpha - \theta$ (assuming that $\theta < \alpha$). Therefore, in our case

$$\tilde{f}(x) = \gamma e^{-\gamma x},\ x \geq 0.$$

satisfies the given constraint and is the solution to this problem. Note here that the tail distribution function equals $\exp(-\gamma x)$ and is heavier than $\exp(-\alpha x)$, the original tail distribution of X.

Polynomial-divergence: Now consider the case where the objective corresponds to a polynomial-divergence with parameter equal to β, i.e, it equals

$$\int \left(\frac{\tilde{f}(x)}{f(x)} \right)^{\beta+1} f(x)dx.$$

Under this objective, the optimal pdf

$$\tilde{f}(x) = \frac{(1+\beta\lambda x)^{1/\beta} \alpha e^{-\alpha x}}{\int (1+\beta\lambda x)^{1/\beta} \alpha e^{-\alpha x}\, dx}$$

where $\lambda > 0$ is chosen so that the mean under $\tilde{f}$ equals $1/\gamma$.

While this may not have a closed form solution, it is clear that on a logarithmic scale, $\tilde{f}(x)$ is asymptotically similar to $\exp(-\alpha x)$ as $x \to \infty$ and hence has a lighter tail than the solution under the KL distance.

Total Variation Distance: Under total variation distance as an objective, we show that given any ε, we can find a new density function $\tilde{f}$ so that the mean under the new distribution equals $1/\gamma$ while the total variation distance is less than ε. Thus the optimal value of the objective function is zero, although there may be no pdf that attains this value.

To see this, consider,

$$\tilde{f}(x) = \varepsilon/2 \frac{I_{(a-\delta, a+\delta)}}{2\delta} + (1 - \varepsilon/2)\alpha e^{-\alpha x},\ x \geq 0.$$

Then,

$$\tilde{E}(X) = \int x\tilde{f}(x)\,dx = (\varepsilon/2)a + \frac{1-\varepsilon/2}{\alpha}.$$

Thus, given any ε if we select

$$a = \frac{1/\gamma - 1/\alpha}{\varepsilon/2} + 1/\alpha$$

we see that

$$\tilde{E}(X) = (\varepsilon/2)a + \frac{1-\varepsilon/2}{\alpha} = 1/\gamma.$$

We now show that total variation distance between f and $\tilde{f}$ is less than ε. To see this, note that

$$|\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq (\varepsilon/2)P(A)$$

for any set A disjoint from $(a-\delta, a+\delta)$, where the probability P corresponds to the density f. Furthermore, letting $L(S)$ denote the Lebesgue measure of set S,

$$|\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq (\varepsilon/4\delta)L(A) + (\varepsilon/2)P(A)$$

for any set $A \subset (a-\delta, a+\delta)$. Thus, for any set $A \subset (0,\infty)$

$$|\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq$$

$$(\varepsilon/4\delta)L(A\bigcap(a-\delta, a+\delta)) + (\varepsilon/2)P(A).$$

Therefore,

$$\sup_A |\int_A f(x)dx - \int_A \tilde{f}(x)dx| \leq (\varepsilon/2)P(A) + (\varepsilon/4\delta)2\delta < \varepsilon.$$

This also illustrates that it may be difficult to have an elegant characterization of solutions under the total variation distance, making the other two as more attractive measures from this viewpoint.

6.2 Views on Probability of Disjoint Sets

Here, we consider the case where the views correspond to probability assignments under posterior measure ν to mutually exclusive and exhaustive set of events and note that objective functions associated with KL distance, polynomial-divergence and total variation distance give identical results.

Suppose that our views correspond to:

$$\nu(B_i) = \alpha_i,\ i = 1, 2, ...k \text{ where}$$

$$B_i's \text{ are disjoint, } \cup B_i = \Omega \text{ and } \sum_{i=1}^{k} \alpha_i = 1.$$

For instance, if L is a continuous random variable denoting loss amount from a portfolio and there is a view that value-at-risk at a certain amount x equals 1%. This may be modeled as $\nu\{L \geq x\} = 1\%$ and $\nu\{L < x\} = 99\%$.

KL Distance: Then, under the KL distance setting, for any event A, the optimal

$$\nu(A) = \frac{\int_A e^{\sum_i \lambda_i I(B_i)}\, d\mu}{\int e^{\sum_i \lambda_i I(B_i)}\, d\mu} = \frac{\sum_i e^{\lambda_i} \mu(A \cap B_i)}{\sum_i e^{\lambda_i} \mu(B_i)}.$$

Select λ_i so that $e^{\lambda_i} = \alpha_i / \mu(B_i)$. Then it follows that the specified views hold and

$$\nu(A) = \sum_i \alpha_i \mu(A \cap B_i)/\mu(B_i). \tag{25}$$

Polynomial-divergence: The analysis remains identical when we use polynomial-divergence with parameter β. Here, we see that optimal

$$\nu(A) = \frac{\sum_i (1+\beta\lambda_i)^{1/\beta} \mu(A \cap B_i)}{\sum_i (1+\beta\lambda_i)^{1/\beta} \mu(B_i)}$$

Again, by setting $(1+\beta\lambda_i)^{1/\beta} = \alpha_i/\mu(B_i)$, (25) holds.

Total Variation Distance: If the objective is the total variation distance, then clearly, the objective function is $\geq \max_i |\mu(B_i) - \alpha_i|$. We now show that ν defined by (25) achieves this lower bound.

To see this, note that

$$\begin{aligned} |\nu(A) - \mu(A)| &= \left| \sum_i (\nu(A \cap B_i) - \mu(A \cap B_i)) \right| \\ &\leq \sum_i |\nu(A \cap B_i) - \mu(A \cap B_i)| \\ &\leq \sum_i \frac{\mu(A \cap B_i)}{\mu(B_i)} |\alpha_i - \mu(B_i)| \\ &\leq \max_i |\mu(B_i) - \alpha_i|. \end{aligned}$$

7 Numerical Experiments

We conduct three simple experiments. In the first, we consider a simple calibration problem, where the distribution of the underlying Black-Scholes model is updated through polynomial divergence to match the observed options prices. We then consider portfolio modeling problem in Markowitz framework, where there is a view that a simple linear combination of securities has a t distribution, along with a view on a moment of another security. In the third example, we empirically observe the parameter space where Assumption 2 holds, in a simple two random variable, two constraint setting.

Example 6 Consider a security whose current price S_0 equals 50. Its volatility σ is estimated to equal 0.2. Suppose that the interest rate r is constant and equals 5%. Consider two liquid European call options on this security with maturity $T = 5$ years, strike prices, respectively, $K_1 = 55$, and $K_2 = 60$, and respective market prices of 15.0000 and 13.0000. It is easily checked that the Black Scholes price of these options at $\sigma = 0.2$ equals 12.2731 and 10.2895, respectively. It can also be easily checked that there is no value of σ making two of the Black-Scholes prices match the observed market prices.

We apply polynomial divergence methodology to arrive at a probability measure closest to the Black-Scholes measure while matching the observed market prices of the two liquid options. Note that under Black-Scholes

$$S(T) \sim \text{ LogNormal}\left(\log S(0) + (r - \frac{\sigma^2}{2})T, \sigma^2 T\right) = \text{ LogNormal}\left(\log 50 + 0.15, 0.2\right)$$

which is heavy-tailed in that the moment generating function does not exist for positive arguments. Let f denote the pdf for LogNormal$(\log 50 + 0.15, 0.2)$. We apply Theorem 2 with $\beta = \frac{1}{n}$ to obtain the posterior distribution:

$$\tilde{f}(x) = \frac{\left(1 + \frac{\lambda_1}{n}(x - K_1)^+ + \frac{\lambda_2}{n}\lambda_2(x - K_2)^+\right)^n f(x)}{\int_0^\infty \left(1 + \frac{\lambda_1}{n}(x - K_1)^+ + \frac{\lambda_2}{n}(x - K_2)^+\right)^n f(x)\, dx}$$

where λ_1 and λ_2 is solved from the constrained equations.

$$\tilde{E}[e^{-rT}(S(T) - K_1)^+] = 15 \quad\text{and}\quad \tilde{E}[e^{-rT}(S(T) - K_2)^+] = 13 \tag{26}$$

with the intention of incorporating these views into our prior model.

Table 1 shows the resulting λ_1 and λ_2 when $\beta = \frac{1}{n},\ n = 1, 2, ..., 7$ (found using subroutine FindRoot of Mathematica). Note that for each n, $\lambda_1 > -\frac{n}{K_2 - K_1}$ and $\lambda_1 + \lambda_2 > 0$. Therefore the tabulated values are all feasible. Furthermore, in each case $\frac{\lambda_1}{n} \times 15 + \frac{\lambda_2}{n} \times 13 > 0$. Therefore they are strongly feasible as well. Plugging these values in $\tilde{f}(\cdot)$ we get the posterior density that can be used to price other options of the same maturity. Table 2 shows the resulting European call option prices for different values of strike prices. Interestingly, changing n affects the resulting prices negligibly except for strongly out of the money options.

n	1	2	3	4	5	6	7
λ_1	-0.107882	-0.1228836	-0.1282311	-0.130952	-0.1325945	-0.133692	-0.1344763
λ_2	0.114	0.1297726	0.135363	0.1381988	0.1399075	0.1410474	0.1418613

Table 1: λ_1 and λ_2 solved from the constrained equation (26) (for $\beta = \frac{1}{n},\ n = 1, 2, ..., 7$).

Example 7 We now consider a small portfolio modeling example involving two assets A_1 and A_2. We assume that the prior distribution of returns (Z_1, Z_2) from assets (A_1, A_2) is bi-variate Gaussian. Specifically,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 9.1 & 3.0 \\ 3.0 & 1.1 \end{bmatrix}\right)$$

Suppose the portfolio management team has the following views on these securities:

- A bench mark portfolio consisting of 70% in A_1 and 30% in A_2 is expected to generate 1.5% average return, while having much heavier tail compared to a Gaussian distribution. This may be modeled as a t-distribution with 3 degrees of freedom and mean= 1.5%.

- Security A_2 will generate 1.5% average return.

Let $X = 0.7Z_1 + 0.3Z_2$, and $Y = Z_2$. Then the above views correspond to

- X_1 has a density function given by $g(x) = \frac{2}{2.4120 \times \pi\sqrt{3}[1 + \frac{1}{3}(\frac{x - 1.5}{2.4120})^2]^2}$.

SP	BS	n=1	n=2	n=3	n=4	n=5	n=6	n=7
50	14.5693	17.4548	17.4528	17.4523	17.4519	17.4519	17.4519	17.4519
51	14.0842	16.9226	16.9210	16.9205	16.9202	16.9203	16.9201	16.9202
52	13.6121	16.4112	16.4100	16.4096	16.4094	16.4094	16.4093	16.4094
53	13.1530	15.9204	15.9196	15.9193	15.9192	15.9193	15.9192	15.9192
54	12.7067	15.4501	15.4497	15.4495	15.4495	15.4496	15.4494	15.4496
55	12.2731	15.0000	15.0000	15.0000	15.0000	15.0001	15.0000	15.0002
56	11.8520	14.5695	14.5699	14.5700	14.5700	14.5702	14.5701	14.5703
57	11.4433	14.1566	14.1572	14.1574	14.1573	14.1575	14.1575	14.1577
58	11.0468	13.7589	13.7594	13.7596	13.7595	13.7597	13.7597	13.7599
59	10.6623	13.3741	13.3743	13.3744	13.3744	13.3745	13.3745	13.3746
60	10.2895	13.0000	13.0000	13.0000	13.0000	13.0001	13.0001	13.0002
61	9.9283	12.6347	12.6347	12.6348	12.6347	12.6349	12.6349	12.6350
62	9.5783	12.2778	12.2782	12.2785	12.2786	12.2787	12.2788	12.2789
63	9.2395	11.9293	11.9304	11.9308	11.9310	11.9314	11.9314	11.9316
64	8.9116	11.5890	11.5911	11.5919	11.5922	11.5927	11.5928	11.5930
65	8.5942	11.2570	11.2603	11.2615	11.2620	11.2626	11.2628	11.2631
70	7.1577	9.7162	9.7283	9.7326	9.7347	9.7362	9.7371	9.7379
75	5.9487	8.3642	8.3880	8.3964	8.4007	8.4034	8.4052	8.4066
80	4.9366	7.1853	7.2215	7.2343	7.2410	7.2452	7.2480	7.2501
85	4.0928	6.1624	6.2105	6.2277	6.2367	6.2424	6.2461	6.2490
90	3.3915	5.2786	5.3371	5.3584	5.3695	5.3766	5.3813	5.3848

Table 2: Option prices for different strikes as computed by the posterior distributions for different n. Here, SP stands for strike price and BS stands for Black Scholes price at $\sigma = 0.2$.

- $\tilde{E}[Y] = 1.5$.

Since under the prior distribution we have:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \\ 0.0 & 1.0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 5.818 & 2.43 \\ 2.43 & 1.1 \end{bmatrix} \right)$$

Therefore we see that $\sigma_{xx} = 5.818$, $\sigma_{xy} = 2.43$ and $\sigma_{yy} = 1.1$ so that

$$\sigma_{yy} - \frac{1}{\sigma_{xx}} \sigma_{xy} \sigma_{xy}^t = 0.08506$$

In this case $a = 1.5$. Therefore

$$a + \left(\frac{x - E_g(X)}{\sigma_{xx}} \right) \sigma_{xy} = 1.5 + \frac{(x - 1.5)}{5.818} \times 2.43 = 0.8735 + 0.41767x$$

By Proposition 3, the posterior distribution of (X, Y) is given by

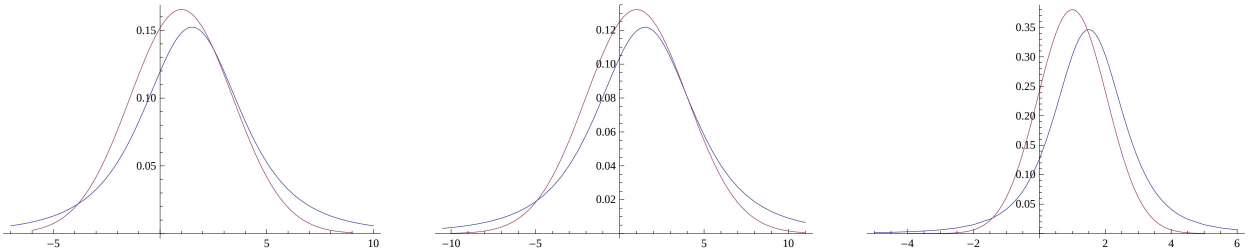
$$\frac{3.42876}{\sqrt{2\pi}} \exp\{-5.8782(y - 0.41767x - 0.8735)^2\} \times \frac{2}{2.4120 \times \pi\sqrt{3}[1 + \frac{1}{3}(\frac{x-1.5}{2.4120})^2]^2}$$

The posterior distribution of (Z_1, Z_2) is given by

$$\tilde{f}(z_1, z_2) = \frac{3.42876 \times 0.7}{\sqrt{2\pi}} \exp\{-5.8782(0.29237z_1 - 0.8747z_2 + 0.8735)^2\}$$

$$\times \frac{2}{2.4120 \times \pi\sqrt{3}[1+\frac{1}{3}(\frac{0.7z_1+0.3z_2-1.5}{2.4120})^2]^2}$$

In Figure 1 we compare the marginal densities of X, Z_1 and Z_2 under prior and posterior distributions.TWe note that incorporating the constraint that X has a fat-tailed density renders the asset returns from A_1 and A_2 to be similarly fat-tailed.



(a) Prior marginal density of X is normal with mean=1 and variance=5.818. Posterior density is t with df=3,mean=1.5 and scale=$\sqrt{5.818} = 2.412$

(b) Prior marginal density of Z_1 is normal with mean=1 and variance=9.1. Posterior density has mean (and mode) 1.5 and heavier tails .

(c) Prior marginal density of Z_2 is normal with mean=1 and variance=1.1. Posterior density has mean (and mode) 1.5 and heavier tails.

Figure 1: Prior and posterior marginal densities under a constraint on the marginal density of a portfolio.

Example 8 In this example we further refine the observation made in Proposition 2 and the following examples that typically the solution space where Assumption 2 holds increases with increasing $n = \frac{1}{\beta}$. We note that even in simple cases, this need not always be true.

Specifically, consider random variables X and Y such that

$$\begin{bmatrix} \log X \\ \log Y \end{bmatrix} \sim \text{Bivariate Gaussian} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

Then, X and Y are Lognormally distributed and their joint density function of (X, Y) is given by:

$$\frac{1}{2\pi xy\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \{(\log x)^2 + (\log y)^2 - 2\rho(\log x)(\log y)\} \right].$$

Consider the constraints,

$$\begin{bmatrix} \tilde{E}(X) \\ \tilde{E}(Y) \end{bmatrix} = \begin{bmatrix} aE(X) \\ bE(Y) \end{bmatrix}$$

Our goal is to find values of a,b for which the associated optimization problem $\mathbf{O_2}(\beta)$ has a solution. The probability distribution minimizing the polynomial divergence with $\beta = \frac{1}{n}$ is of the form:

$$\tilde{f}(x, y) = \frac{(1 + \frac{\lambda}{n}x + \frac{\xi}{n}y)^n f(x, y)}{\int_0^\infty (1 + \frac{\lambda}{n}x + \frac{\xi}{n}y)^n f(x, y)\, dxdy} = \frac{(1 + \frac{\lambda}{n}x + \frac{\xi}{n}y)^n f(x, y)}{E[(1 + \frac{\lambda}{n}X + \frac{\xi}{n}Y)^n]}.$$

Now from the constraint equations we have

$$a = \frac{E[X(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}{E[X]E[(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}$$

$$b = \frac{E[Y(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}{E[Y]E[(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}$$

Note that since X and Y takes values in $(0,\infty)$ only $\lambda \geq 0$ and $\xi \geq 0$ are feasible. Using ParametricPlot of Mathematica we plot the values of $(\frac{E[X(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}{E[X]E[(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}, \frac{E[Y(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]}{E[Y]E[(1+\frac{\lambda}{n}X+\frac{\xi}{n}Y)^n]})$ for λ and ξ in the range $[0,10n]$. Figure (2), depicts the range when $\rho = -\frac{1}{4}$, $\rho = 0$, $\rho = \frac{1}{4}$ and $\rho = \frac{1}{2}$ respectively, for $n = 1, n = 2$ and $n = 3$.

From the graph it appears that the solution space strictly increases with n when $\rho \geq 0$. However, this is not true for $\rho < 0$.

8 Conclusion

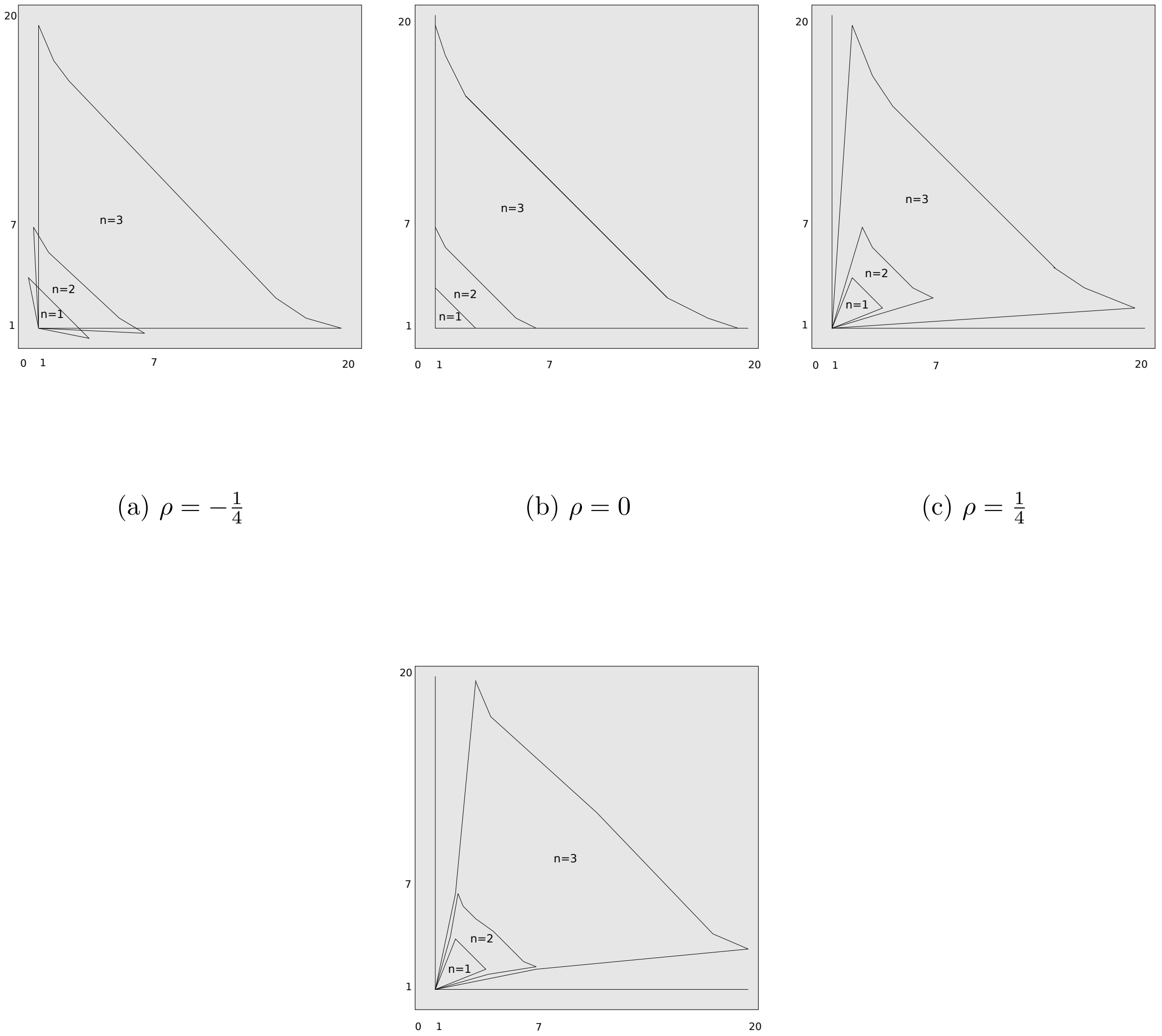
In this article, we built upon existing methodologies that use KL entropy based ideas for incorporating mathematically specified views/constraints to a given financial model to arrive at a more accurate one, when the underlying random variables are light tailed. In the existing financial literature, these ideas have found applications in portfolio modeling and in model calibration. Our key contribution was to show that under technical conditions, using polynomial divergence, such constraints may be uniquely incorporated even when the underlying random variables have fat tails. We also extended the proposed methodology to allow for constraints on marginal distributions on functions of underlying variables. This, in addition to the constraints on moments of functions of underlying random variables, traditionally considered for such problems. Here, we considered, both KL distance and polynomial-divergence as objective. Both these proposals maybe useful for effective risk management.

We also specialized our results to the Markowitz portfolio modeling settings where multivariate Gaussian distribution is used to model asset returns. Here, we developed close form solutions for the updated posterior distribution. In case when there is a constraint that a marginal of a single portfolio of assets has a heavy tailed distribution, we showed that under the posterior distribution, marginal of all assets with non-zero correlation this portfolio have similar fat-tailed distribution. This may be a reasonable and a simple way to incorporate realistic tail behavior in a portfolio of assets.

We also qualitatively compared the solution to the optimization problem in a simple setting of exponentially distributed prior when the objective function was KL distance, polynomial-divergence and total variational distance. We found that in certain settings, KL distance may put more mass in tails compared to polynomial-divergence, which may penalize tail deviation more. Finally, we numerically tested the proposed methodology on some simple examples.

While we show that the proposed solutions when they exist are unique under mild conditions, we do not give conditions under which existence of the proposed results is guaranteed. This remains an interesting area for further research.

Acknowledgement: *The authors would like to thank Paul Glasserman for many directional suggestions as well as specific inputs on the manuscript that greatly helped this effort.*



(a) $\rho = -\frac{1}{4}$

(b) $\rho = 0$

(c) $\rho = \frac{1}{4}$

(d) $\rho = \frac{1}{2}$

Figure 2: Solution range as a function of correlation and n.

9 Appendix: Proofs

Proof of Proposition 1 Recall that $f(x) = \frac{\alpha-1}{(1+x)^\alpha}$ for $x \geq 0$, and

$$f_{\lambda_M}(x) = \frac{e^{\lambda_M x} f(x) I_{[0,M]}(x)}{\int_0^M e^{\lambda_M x} f(x) dx}$$

where λ_M is the unique solution to $\int_0^M x f_\lambda(x)\,dx = c$.

We first show that $\lambda_M \to 0$ as $M \to \infty$. To this end, let $g(\lambda, M) = \frac{\int_0^M x e^{\lambda x} f(x)\,dx}{\int_0^M e^{\lambda x} f(x)\,dx}$, $M \geq 1$, $\lambda \geq 0$.

We have

$$\begin{aligned}\frac{\partial g}{\partial \lambda} &= \frac{\left(\int_0^M x^2 e^{\lambda x} f(x)\,dx\right)\left(\int_0^M e^{\lambda x} f(x)\,dx\right) - \left(\int_0^M x e^{\lambda x} f(x)\,dx\right)^2}{\left(\int_0^M e^{\lambda x} f(x)\,dx\right)^2} \\ &= \int_0^M x^2 \left(\frac{e^{\lambda x} f(x)}{\int_0^M e^{\lambda x} f(x)\,dx}\right) dx - \left(\int_0^M x \left(\frac{e^{\lambda x} f(x)}{\int_0^M e^{\lambda x} f(x)\,dx}\right) dx\right)^2 \\ &= E_\lambda[X^2] - (E_\lambda[X])^2 > 0\end{aligned}$$

where E_λ is the expectation operator w.r.t density function f_λ.

Also

$$\begin{aligned}\frac{\partial g}{\partial M} &= \frac{\left(Me^{\lambda M} f(M)\right)\left(\int_0^M e^{\lambda x} f(x)\,dx\right) - \left(\int_0^M x e^{\lambda x} f(x)\,dx\right)\left(e^{\lambda M} f(M) - f(0)\right)}{\left(\int_0^M e^{\lambda x} f(x)\,dx\right)^2} \\ &= \frac{e^{\lambda M} f(M)\int_0^M (M-x) e^{\lambda x} f(x)\,dx + f(0)\int_0^M e^{\lambda x} f(x)\,dx}{\left(\int_0^M e^{\lambda x} f(x)\,dx\right)^2} > 0.\end{aligned}$$

Since λ_M satisfies $g(\lambda_M, M) = c$, it follows that increasing M leads to reduction in λ_M. That is, λ_M is a non-increasing function of M.

Suppose that $\lambda_M \downarrow c_1 > 0$. Then $c = g(\lambda_M, M) \geq g(c_1, M)$ for all M. But since $c_1 > 0$ we have $g(c_1, M) \to \infty$ as $M \to \infty$, a contradiction. Hence, $\lambda_M \to 0$ as $M \to \infty$.

Next, since

$$\int_0^M \log\left(\frac{f_\lambda(x)}{f(x)}\right) f_\lambda(x)\,dx = \lambda \int_0^M x f_\lambda(x)\,dx - \log\left(\int_0^M e^{\lambda x} f(x)\,dx\right)$$

we see that

$$\int_0^M \log\left(\frac{f_{\lambda_M}(x)}{f(x)}\right) f_{\lambda_M}(x)\,dx = \lambda_M c - \log\left(\int_0^M e^{\lambda_M x} f(x)\,dx\right)$$

Hence, to prove that the LHS converges to zero as $M \to \infty$, it suffices to show that $\int_0^M e^{\lambda_M x} f(x)\,dx \to 1$ or that $\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx \to \frac{1}{\alpha-1}$.

Note that the constraint equation can be re-expressed as:

$$c = \int_0^M x f_{\lambda_M}(x)\,dx = \frac{\int_0^M \frac{x e^{\lambda_M x}}{(1+x)^\alpha}\,dx}{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx} = \frac{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^{\alpha-1}}\,dx - \int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx}{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx} = \frac{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^{\alpha-1}}\,dx}{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx} - 1,$$

or

$$\frac{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^{\alpha-1}}\,dx}{\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx} = 1 + c. \tag{27}$$

Further, by integration by parts of the numerator, we observe that

$$\int_0^M \frac{e^{\lambda_M x}}{(1+x)^{\alpha-1}}\,dx = \frac{1}{\lambda_M}\left(\frac{e^{\lambda_M M}}{(1+M)^{\alpha-1}} - 1 + (\alpha-1)\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx\right).$$

From the above equation and (27), it follows that

$$\int_0^M \frac{e^{\lambda_M x}}{(1+x)^\alpha}\,dx = \frac{\frac{e^{\lambda_M M}}{(1+M)^{\alpha-1}} - 1}{\lambda_M(c+1) - (\alpha-1)} \tag{28}$$

Since $\lambda_M \to 0$, it suffices to show that

$$\frac{e^{\lambda_M M}}{(1+M)^{\alpha-1}} \to 0.$$

Suppose this is not true. Then, there exists an $\eta > 0$ and a sequence $M_i \uparrow \infty$ such that $\frac{e^{\lambda_{M_i} M_i}}{(1+M_i)^{\alpha-1}} \geq \eta$.

Equation (27) may be re-expressed as:

$$\int_0^M \frac{e^{\lambda_M x}}{(1+x)^{\alpha-1}} \left[1 - \frac{1+c}{1+x}\right] dx = 0 \tag{29}$$

Given an arbitrary $K > 0$, one can find an $M_i \geq 1+2c$ (so that $1 - \frac{1+c}{1+x} \geq \frac{1}{2}$ when $x \geq M_i$) such that, for $x \in [M_i - K, M_i]$

$$\frac{e^{\lambda_{M_i} x}}{(1+x)^{\alpha-1}} \geq \frac{e^{\lambda_{M_i}(M_i-K)}}{(1+M_i)^{\alpha-1}} \geq \frac{\eta}{2}.$$

Re-expressing the LHS of (29) evaluated at $M = M_i$ as

$$\int_0^c \frac{e^{\lambda_{M_i} x}}{(1+x)^{\alpha-1}} \left[1 - \frac{1+c}{1+x}\right] dx + \int_c^{M_i-K} \frac{e^{\lambda_{M_i} x}}{(1+x)^{\alpha-1}} \left[1 - \frac{1+c}{1+x}\right] dx + \int_{M_i-K}^{M_i} \frac{e^{\lambda_{M_i} x}}{(1+x)^{\alpha-1}} \left[1 - \frac{1+c}{1+x}\right] dx,$$

we see that this is bounded from below by

$$-c^2 e^{\lambda_{M_i} c} + K\eta/4.$$

For sufficiently large K, this is greater than zero providing the desired contradiction to (29).□

Proof of Theorem 2:

Let $\xi = \int \left(1 + \beta \sum_l \lambda_l g_l\right)^{1/\beta} d\mu$. and $\hat{\lambda}_l = (\beta+1)\beta\lambda_l/\xi^\beta$. Consider the Lagrangian $\mathcal{L}(\nu)$ for $\mathbf{O_2}(\beta)$ defined as

$$\int \left(\frac{d\nu}{d\mu}\right)^{\beta+1} d\mu - \sum_l \hat{\lambda}_l \left(\int g_l\, d\nu - c_l\right). \tag{30}$$

We first argue that $\mathcal{L}(\nu)$ is a convex function of ν. Given that $\lambda_l \int g_l\, d\nu$ are linear in ν, it suffices to show that $\int (\frac{d\nu}{d\mu})^{\beta+1}\, d\mu$ is a convex function of ν.

Note that for $0 \leq s \leq 1$,

$$\int \left(\frac{d(s\nu_1 + (1-s)\nu_2)}{d\mu}\right)^{\beta+1} d\mu$$

equals

$$\int \left(s\frac{d\nu_1}{d\mu} + (1-s)\frac{d\nu_2}{d\mu}\right)^{\beta+1} d\mu$$

which in turn is dominated by

$$\int \left[s \left(\frac{d\nu_1}{d\mu} \right)^{\beta+1} + (1-s) \left(\frac{d\nu_2}{d\mu} \right)^{\beta+1} \right] d\mu$$

which equals

$$s \int \left(\frac{d\nu_1}{d\mu} \right)^{\beta+1} d\mu + (1-s) \int \left(\frac{d\nu_2}{d\mu} \right)^{\beta+1} d\mu.$$

Therefore, the Lagrangian $\mathcal{L}(\nu)$ is a convex function of ν.

We now prove that $\mathcal{L}(\nu)$ is minimized at ν^1. For this, all we need to show is that we cannot improve by moving in any feasible direction away from ν^1. Since, ν^1 satisfies all the constraints, the result then follows. We now show this.

Let f denote $\frac{d\nu}{d\mu}$ and $f^1 = \frac{d\nu^1}{d\mu}$. Note that (30) may be re-expressed as

$$\int \left(f^{\beta+1} - \sum_l \hat{\lambda}_l g_l f \right) d\mu + \sum_l \hat{\lambda}_l c_l.$$

For any $\nu \in \mathcal{P}(\mu)$ and $t \in [0,1]$ consider the function

$$G_\nu(t) = \mathcal{L}\left((1-t)\nu^1 + t\nu \right).$$

This in turn equals

$$\int \left[\{(1-t)f^1 + tf\}^{\beta+1} - \sum_l \hat{\lambda}_l g_l \{(1-t)f^1 + tf\} \right] d\mu + \sum_l \hat{\lambda}_l c_l.$$

We now argue that $\frac{d}{dt}_{t=0} G_\nu(t) = 0$. Then from this and convexity of $\mathcal{L}$, the result follows.

To see this, note that $\frac{d}{dt}_{t=0} G_\nu(t)$ equals

$$\int \left\{ (\beta+1)(f^1)^\beta - \sum_l \hat{\lambda}_l g_l \right\} (f - f^1)\, d\mu. \tag{31}$$

Due to the definition of f_1 and $\hat{\lambda}_i$, it follows that the term inside the braces in the integrand in (31) is a constant. Since both ν^1 and ν are probability measures, therefore $\frac{d}{dt}_{t=0} G_\nu(t) = 0$ and the result follows.□

Proof of Theorem 3: Let A denote the set of all strongly feasible solutions to (11). Consider the following set of equations:

$$\frac{\int g_j \left(1 + \beta \sum_{l=1}^k \theta_l (g_l - c_l) \right)^{\frac{1}{\beta}} d\mu}{\int \left(1 + \beta \sum_{l=1}^k \theta_l (g_l - c_l) \right)^{\frac{1}{\beta}} d\mu} = c_j \ \text{ for } j = 1, 2, \ldots, k. \tag{32}$$

We say that $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ is a *strongly feasible* solution to (32) if it solves (32), and lies in the set:

$$\{\boldsymbol{\theta} \in \mathbb{R}^k \mid 1 + \beta \sum_{l=1}^k \theta_l (g_l - \theta_l) \geq 0 \ a.e.(\mu),$$

$$\int (1+\beta\sum_{l=1}^{k}\theta_l(g_l-c_l))^{\frac{1}{\beta}+1}d\mu < \infty \text{ and } 1-\beta\sum_{l=1}^{k}\theta_l c_l > 0\}.$$

Let B denote the set of all strongly feasible solutions to (32).

Let $\phi : A \to B$ and $\psi : B \to A$ be the mappings defined by

$$\phi(\boldsymbol{\lambda}) = \left(\frac{\lambda_1}{1+\beta\sum_{l=1}^{k}\lambda_l c_l}, \frac{\lambda_2}{1+\beta\sum_{l=1}^{k}\lambda_l c_l}, \ldots, \frac{\lambda_k}{1+\beta\sum_{l=1}^{k}\lambda_l c_l}\right)$$

and

$$\psi(\boldsymbol{\theta}) = \left(\frac{\theta_1}{1-\beta\sum_{l=1}^{k}\theta_l c_l}, \frac{\theta_2}{1-\beta\sum_{l=1}^{k}\theta_l c_l}, \ldots, \frac{\theta_k}{1-\beta\sum_{l=1}^{k}\theta_l c_l}\right).$$

It is easily checked that mapping ϕ is a bijection with inverse ψ. Note that if $\boldsymbol{\lambda} \in A$, then $\phi(\boldsymbol{\lambda}) \in B$. To see this, simply divide the numerator and the denominator in (11) by $(1+\beta\sum_{l=1}^{k}\lambda_l c_l)^{1/\beta}$. Conversely, if $\boldsymbol{\theta} \in B$, then $\psi(\boldsymbol{\theta}) \in A$. To see this, divide the numerator and the denominator in (32) by $(1-\beta\sum_{l=1}^{k}\theta_l c_l)^{1/\beta}$.

Therefore, its suffices to prove uniqueness of strongly feasible solutions to (32). To this end, consider the function $G : B \to \mathbb{R}^+$:

$$G(\boldsymbol{\theta}) = \int \left(1+\beta\sum_{l=1}^{k}\theta_l(g_l-c_l)\right)^{\frac{1}{\beta}+1} d\mu.$$

Then,

$$\frac{\partial G}{\partial \theta_i} = (1+\beta)\int (g_i-c_i)\left(1+\beta\sum_{l=1}^{k}\theta_l(g_l-c_l)\right)^{\frac{1}{\beta}} d\mu. \tag{33}$$

and

$$\frac{\partial^2 G}{\partial \theta_j \partial \theta_i} = \beta(1+\beta)\int (g_i-c_i)(g_j-c_j)\left(1+\beta\sum_{l=1}^{k}\theta_l(g_l-c_l)\right)^{\frac{1}{\beta}-1} d\mu\,.$$

We see that the last integral can be written as $E_\theta[(g_i-c_i)(g_j-c_j)]$ times a positive constant independent of i and j, where E_θ denotes expectation under the measure

$$\frac{\left(1+\beta\sum_{l=1}^{k}\theta_l(g_l-c_l)\right)^{\frac{1}{\beta}-1} d\mu}{\int \left(1+\beta\sum_{l=1}^{k}\theta_l(g_l-c_l)\right)^{\frac{1}{\beta}-1} d\mu}.$$

Now from the identity

$$E_\theta[(g_i-c_i)(g_j-c_j)] = \mathrm{Cov}_\theta[g_i,g_j] + (E_\theta[g_i]-c_i)\,(E_\theta[g_j]-c_j)$$

and the assumption on g_i's it follows that the Hessian of the function G is positive definite. Thus, G is strictly convex in its domain of definition, that is in B. Therefore if a solution to the equation

$$\left(\frac{\partial G}{\partial \theta_1}, \frac{\partial G}{\partial \theta_2}, \ldots, \frac{\partial G}{\partial \theta_k}\right) = 0 \tag{34}$$

exists in B, then it is unique. From (33) it follows that the set of equations given by (34) and (32) are equivalent. □

Proof of Theorem 4: In view of (14), we may fix the marginal distribution of $\mathbf{X}$ to be $g(\mathbf{x})$ and re-express the objective as

$$\min_{\tilde{f}(\cdot|\mathbf{x})\in\mathcal{P}(f(\cdot|\mathbf{x})),\forall\mathbf{x}} \int_{\mathbf{x},\mathbf{y}} \log\left(\frac{\tilde{f}(\mathbf{y}|\mathbf{x})}{f(\mathbf{y}|\mathbf{x})}\right)\tilde{f}(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{y}d\mathbf{x} + \int_{\mathbf{x}}\log\left(\frac{g(\mathbf{x})}{f(\mathbf{x})}\right)g(\mathbf{x})d\mathbf{x}\,.$$

The second integral is a constant and can be dropped from the objective. The first integral may in turn be expressed as

$$\int_{\mathbf{x}}\min_{\tilde{f}(\cdot|\mathbf{x})\in\mathcal{P}(f(\cdot|\mathbf{x}))}\left(\int_{\mathbf{y}}\log\frac{\tilde{f}(\mathbf{y}|\mathbf{x})}{f(\mathbf{y}|\mathbf{x})}\tilde{f}(\mathbf{y}|\mathbf{x})d\mathbf{y}\right)g(\mathbf{x})d\mathbf{x}\,.$$

Similarly the moment constraints can be re-expressed as

$$\int_{\mathbf{x},\mathbf{y}} h_i(\mathbf{x},\mathbf{y})\tilde{f}(\mathbf{y}|\mathbf{x})g(\mathbf{x})d\mathbf{x}d\mathbf{y} = c_i,\;\; i=1,2,...,k$$

or

$$\int_{\mathbf{x}}\left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y})\tilde{f}(\mathbf{y}|\mathbf{x})d\mathbf{y}\right)g(\mathbf{x})d\mathbf{x} = c_i,\;\; i=1,2,...,k\,.$$

Then, the Lagrangian for this k constraint problem is,

$$\int_{\mathbf{x}}\left[\min_{\tilde{f}(\cdot|\mathbf{x})\in\mathcal{P}(f(\cdot|\mathbf{x}))}\int_{\mathbf{y}}\left(\log\frac{\tilde{f}(\mathbf{y}|\mathbf{x})}{f(\mathbf{y}|\mathbf{x})}\tilde{f}(\mathbf{y}|\mathbf{x}) - \sum_i \delta_i h_i(\mathbf{x},\mathbf{y})\tilde{f}(\mathbf{y}|\mathbf{x})\right)d\mathbf{y}\right]g(\mathbf{x})d\mathbf{x} + \sum_i \delta_i c_i\,.$$

Note that by Theorem (1)

$$\min_{\tilde{f}(\cdot|\mathbf{x})\in\mathcal{P}(f(\cdot|\mathbf{x}))}\int_{\mathbf{y}}\left(\log\frac{\tilde{f}(\mathbf{y}|\mathbf{x})}{f(\mathbf{y}|\mathbf{x})}\tilde{f}(\mathbf{y}|\mathbf{x}) - \sum_i \delta_i h_i(\mathbf{x},\mathbf{y})\tilde{f}(\mathbf{y}|\mathbf{x})\right)dy$$

has the solution

$$\tilde{f}_{\boldsymbol{\delta}}(\mathbf{y}|\mathbf{x}) = \frac{\exp(\sum_i \delta_i h_i(\mathbf{x},\mathbf{y}))f(\mathbf{y}|\mathbf{x})}{\int_{\mathbf{y}}\exp(\sum_i \delta_i h_i(\mathbf{x},\mathbf{y}))f(\mathbf{y}|\mathbf{x})d\mathbf{y}} = \frac{\exp(\sum_i \delta_i h_i(\mathbf{x},\mathbf{y}))f(\mathbf{x},\mathbf{y})}{\int_{\mathbf{y}}\exp(\sum_i \delta_i h_i(\mathbf{x},\mathbf{y}))f(\mathbf{x},\mathbf{y})d\mathbf{y}}\,.$$

Now taking $\boldsymbol{\delta} = \boldsymbol{\lambda}$, it follows from Assumption (3) that $f_{\boldsymbol{\lambda}}(\mathbf{x},\mathbf{y}) = \tilde{f}_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x})g(\mathbf{x})$ is a solution to $\mathbf{O_3}$.□

Proof of Theorem 5: Let $F:\mathbb{R}^k\to\mathbb{R}$ be a function defined as

$$F(\boldsymbol{\lambda}) = \int_{\mathbf{x}}\log\left(\int_{\mathbf{y}} exp\left(\sum_l \lambda_l h_l(\mathbf{x},\mathbf{y})\right)f(\mathbf{y}|\mathbf{x})d\mathbf{y}\right)g(\mathbf{x})d\mathbf{x} - \sum_l \lambda_l c_l.$$

Then,

$$\begin{aligned}
\frac{\partial F}{\partial \lambda_i} &= \int_{\mathbf{x}} \left(\frac{\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) exp\left(\sum_l \lambda_l h_l(\mathbf{x},\mathbf{y})\right) f(\mathbf{y}|\mathbf{x}) d\mathbf{y}}{\int_{\mathbf{y}} exp\left(\sum_l \lambda_l h_l(\mathbf{x},\mathbf{y})\right) f(\mathbf{y}|\mathbf{x}) d\mathbf{y}} \right) g(\mathbf{x}) d\mathbf{x} - c_i \\
&= \int_{\mathbf{x}} \left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) \frac{exp\left(\sum_l \lambda_l h_l(\mathbf{x},\mathbf{y})\right) f(\mathbf{y}|\mathbf{x})}{\int_{\mathbf{y}} exp\left(\sum_l \lambda_l h_l(\mathbf{x},\mathbf{y})\right) f(\mathbf{y}|\mathbf{x}) d\mathbf{y}} d\mathbf{y} \right) g(\mathbf{x}) d\mathbf{x} - c_i \\
&= \int_{\mathbf{x}} \left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) g(\mathbf{x}) d\mathbf{x} - c_i \\
&= \int_{\mathbf{x}} \int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{x},\mathbf{y}) d\mathbf{x} d\mathbf{y} - c_i \\
&= E_{\boldsymbol{\lambda}}[h_i(\mathbf{X},\mathbf{Y})] - c_i .
\end{aligned}$$

Hence the set of equations given by (16) is equivalent to:

$$\left(\frac{\partial F}{\partial \lambda_1}, \frac{\partial F}{\partial \lambda_2}, \ldots, \frac{\partial F}{\partial \lambda_k} \right) = 0 \,. \tag{35}$$

Since

$$\frac{\partial}{\partial \lambda_j} f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) = h_j(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) - \left(\int_{\mathbf{y}} h_j(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) \times f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}),$$

we have

$$\begin{aligned}
\frac{\partial^2 F}{\partial \lambda_j \partial \lambda_i} &= \int_{\mathbf{x}} \left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) \frac{\partial}{\partial \lambda_j} f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \right) g(\mathbf{x}) \, d\mathbf{x} \\
&= \int_{\mathbf{x}} \left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) h_j(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) g(\mathbf{x}) d\mathbf{x} \\
&\quad - \int_{\mathbf{x}} \left(\int_{\mathbf{y}} h_j(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) \left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) f_{\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) g(\mathbf{x}) d\mathbf{x} \\
&= E_{g(\mathbf{x})} \left[E_{\boldsymbol{\lambda}}[h_i(\mathbf{X},\mathbf{Y}) h_j(\mathbf{X},\mathbf{Y}) \mid \mathbf{X}] \right] - E_{g(\mathbf{x})} \left[E_{\boldsymbol{\lambda}}[h_j(\mathbf{X},\mathbf{Y}) \mid \mathbf{X}] \times E_{\boldsymbol{\lambda}}[h_i(\mathbf{X},\mathbf{Y}) \mid \mathbf{X}] \right] \\
&= E_{g(\mathbf{x})} \left[\mathrm{Cov}_{\boldsymbol{\lambda}}[h_i(\mathbf{X},\mathbf{Y}), h_j(\mathbf{X},\mathbf{Y}) \mid \mathbf{X}] \right]
\end{aligned}$$

Where $E_{g(\mathbf{x})}$ denote expectation with respect to the density function $g(\mathbf{x})$. By our assumption, it follows that the Hessian of F is positive definite. Thus, the function F is strictly convex in $\mathbb{R}^k$. Therefore if there exist a solution to (35), then it is unique. Since (35) is equivalent to (16), the theorem follows.□

Proof of Theorem 6: Fixing the marginal of $\mathbf{X}$ to be $g(\mathbf{x})$ we express the objective as

$$\min_{\tilde{f}(\cdot|\mathbf{x}) \in \mathcal{P}(f(\cdot|\mathbf{x})), \forall \mathbf{x}} \int_{\mathbf{x},\mathbf{y}} \left(\frac{\tilde{f}(\mathbf{y}|\mathbf{x}) g(\mathbf{x})}{f(\mathbf{y}|\mathbf{x}) f(\mathbf{x})} \right)^{\beta} \tilde{f}(\mathbf{y}|\mathbf{x}) g(\mathbf{x}) d\mathbf{y} d\mathbf{x} \,.$$

This may in turn be expressed as

$$\int_{\mathbf{x}} \min_{\tilde{f}(\cdot|\mathbf{x})\in\mathcal{P}(f(\cdot|\mathbf{x}))} \left(\int_{\mathbf{y}} \left(\frac{\tilde{f}(\mathbf{y}|\mathbf{x})}{f(\mathbf{y}|\mathbf{x})} \right)^{\beta} \tilde{f}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) \left(\frac{g(\mathbf{x})}{f(\mathbf{x})} \right)^{\beta} g(\mathbf{x}) d\mathbf{x} \,.$$

Similarly, the moment constraints can be re-expressed as

$$\int_{\mathbf{x}} \left(\int_{\mathbf{y}} h_i(\mathbf{x},\mathbf{y}) \left(\frac{f(\mathbf{x})}{g(\mathbf{x})} \right)^{\beta} \tilde{f}(\mathbf{y}|\mathbf{x}) d\mathbf{y} \right) \left(\frac{g(\mathbf{x})}{f(\mathbf{x})} \right)^{\beta} g(\mathbf{x}) d\mathbf{x} = c_i, \;\; i = 1, 2, ..., k \,.$$

Then, the Lagrangian for this k constraint problem is, up to the constant $\sum_i \delta_i c_i$,

$$\int_{\mathbf{x}} \left[\min_{\tilde{f}(\cdot|\mathbf{x})\in\mathcal{P}(f(\cdot|\mathbf{x}))} \int_{\mathbf{y}} \left(\left(\frac{\tilde{f}(\mathbf{y}|\mathbf{x})}{f(\mathbf{y}|\mathbf{x})} \right)^{\beta} \tilde{f}(\mathbf{y}|\mathbf{x}) - \sum_i \delta_i h_i(\mathbf{x},\mathbf{y}) \left(\frac{f(\mathbf{x})}{g(\mathbf{x})} \right)^{\beta} \tilde{f}(\mathbf{y}|\mathbf{x}) \right) d\mathbf{y} \right] \left(\frac{g(\mathbf{x})}{f(\mathbf{x})} \right)^{\beta} g(\mathbf{x}) d\mathbf{x} \,.$$

By Theorem (2), the inner minimization has the solution $f_{\boldsymbol{\delta},\beta}(\mathbf{y}|\mathbf{x})$. Now taking $\boldsymbol{\delta} = \boldsymbol{\lambda}$, it follows from Assumption (4) that $f_{\boldsymbol{\lambda},\beta}(\mathbf{x},\mathbf{y}) = f_{\boldsymbol{\lambda},\beta}(\mathbf{y}|\mathbf{x}) g(\mathbf{x})$ is the solution to $\mathbf{O_4}(\beta)$. □

Proof of Proposition 2: In view of Assumption (2), we note that $1 + \frac{\lambda}{n} g(x) \geq 0$ for all $x \geq 0$ if $\lambda \geq 0$. By Theorem (2), the probability distribution minimizing the polynomial-divergence (with $\beta = 1/n$) w.r.t. f is given by:

$$\tilde{f}(x) = \frac{\left(1 + \frac{\lambda}{n} g(x)\right)^n f(x)}{c} \;, x \geq 0,$$

where

$$c = \int_0^{\infty} \left(1 + \frac{\lambda}{n} g(x) \right)^n f(x)\, dx = \sum_{k=0}^{n} n^{-k} \binom{n}{k} E[g(X)^k] \lambda^k.$$

From the constraint equation we have

$$a = \frac{\tilde{E}[g(X)]}{E[g(X)]} = \frac{\int_0^{\infty} g(x) \left(1 + \frac{\lambda}{n} g(x)\right)^n f(x) dx}{c E[g(X)]} = \frac{\sum_{k=0}^{n} n^{-k} \binom{n}{k} E[g(X)^{k+1}] \lambda^k}{\sum_{k=0}^{n} n^{-k} \binom{n}{k} E[g(X)] E[g(X)^k] \lambda^k}.$$

Since, $E[g(X)^{n+1}] > a E[g(X)] E[g(X)^n]$, the n-th degree term in (12) is strictly positive and the constant term is negative so there exists a positive λ that solves this equation. Uniqueness of the solution now follows from Theorem 3. □.

Proof of Proposition 3: By Theorem 4:

$$\tilde{f}(\mathbf{x},\mathbf{y}) = g(\mathbf{x}) \times \tilde{f}(\mathbf{y}|\mathbf{x})$$

where

$$\tilde{f}(\mathbf{y}|\mathbf{x}) = \frac{e^{\boldsymbol{\lambda}^t \mathbf{y}} f(\mathbf{y}|\mathbf{x})}{\int e^{\boldsymbol{\lambda}^t \mathbf{y}} f(\mathbf{y}|\mathbf{x}) d\mathbf{y}}.$$

Here the superscript t corresponds to the transpose. Now $f(\mathbf{y}|\mathbf{x})$ is the k-variate normal density with mean vector:

$$\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = \boldsymbol{\mu}_{\mathbf{y}} + \boldsymbol{\Sigma}_{\mathbf{yx}} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})$$

and the variance-covariance matrix:

$$\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}} = \mathbf{\Sigma}_{\mathbf{yy}} - \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{\Sigma}_{\mathbf{xy}}.$$

Hence $\tilde{f}(\mathbf{y}|\mathbf{x})$ is the normal density with mean $(\boldsymbol{\mu}_{\mathbf{y}|x} + \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}\boldsymbol{\lambda})$ and variance-covariance matrix $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}$. Now the moment constraint equation (21) implies:

$$\begin{aligned}
\mathbf{a} &= \int_{\mathbf{x}\in\mathbb{R}^{N-k}} \int_{\mathbf{y}\in\mathbb{R}^{k}} \mathbf{y}\tilde{f}(\mathbf{x},\mathbf{y})d\mathbf{y}d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathbb{R}^{N-k}} g(\mathbf{x}) \left(\int_{\mathbf{y}\in\mathbb{R}^{k}} \mathbf{y}\tilde{f}(\mathbf{y}|\mathbf{x})d\mathbf{y} \right) d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathbb{R}^{N-k}} g(\mathbf{x}) \left(\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} + \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}\boldsymbol{\lambda} \right) d\mathbf{x} \\
&= \int_{\mathbf{x}\in\mathbb{R}^{N-k}} g(\mathbf{x}) \left(\boldsymbol{\mu}_{\mathbf{y}} + \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) + \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}\boldsymbol{\lambda} \right) d\mathbf{x} \\
&= \boldsymbol{\mu}_{\mathbf{y}} + \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}(E_g[\mathbf{X}] - \boldsymbol{\mu}_{\mathbf{x}}) + \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}\boldsymbol{\lambda}.
\end{aligned}$$

Therefore, to satisfy the moment constraint, we must take

$$\boldsymbol{\lambda} = \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}^{-1} \left[\mathbf{a} - \boldsymbol{\mu}_{\mathbf{y}} - \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}(E_g[\mathbf{X}] - \boldsymbol{\mu}_{\mathbf{x}}) \right] .$$

Putting the above value of $\boldsymbol{\lambda}$ in $(\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} + \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}\boldsymbol{\lambda})$ we see that $\tilde{f}(\mathbf{y}|\mathbf{x})$ is the normal density with mean

$$\mathbf{a} + \mathbf{\Sigma}_{\mathbf{yx}}\mathbf{\Sigma}_{\mathbf{xx}}^{-1}(\mathbf{x} - E_g[\mathbf{X}])$$

and variance-covariance matrix $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}$.□

Proof of Theorem 8: We have

$$\tilde{f}(\mathbf{y}|x) = D\exp\{-\frac{1}{2}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_{\mathbf{y}|x})^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_{\mathbf{y}|x})\}$$

for an appropriate constant D, where $\tilde{\boldsymbol{\mu}}_{\mathbf{y}|x}$ denotes $\mathbf{a} + \left(\frac{x - E_g(X)}{\sigma_{xx}}\right)\boldsymbol{\sigma}_{x\mathbf{y}}$.

Suppose that the stated assumptions hold for $i = 1$

Under the optimal distribution, the marginal density of Y_1 is

$$\tilde{f}_{Y_1}(y_1) = \int_{(x,y_2,\dots,y_k)} D\exp\{-\frac{1}{2}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_{\mathbf{y}|x})^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_{\mathbf{y}|x})\}g(x)dxdy_2...dy_k$$

Now the limit in (24) is equal to:

$$\lim_{y_1\to\infty} \int_{(x,y_2,y_3,\dots,y_k)} D\exp\{-\frac{1}{2}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_{\mathbf{y}|x})^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_{\mathbf{y}|x})\} \times \frac{g(x)}{g(y_1)}dxdy_2...dy_k$$

The term in the exponent is:

$$-\frac{1}{2}\sum_{i=1}^{k} \left(\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\right)_{ii} \{(y_i - a_i') - x\frac{\sigma_{xy_i}}{\sigma_{xx}}\}^2 +$$

$$(-\frac{1}{2})\sum_{i\neq j}\left(\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\right)_{ij}\{(y_i-a_i')-x\frac{\sigma_{xy_i}}{\sigma_{xx}}\}\{(y_j-a_j')-x\frac{\sigma_{xy_j}}{\sigma_{xx}}\}$$

where $a_i'=a_i-\frac{E_g(X)}{\sigma_{xx}}\sigma_{xy_i}$.

We make the following substitutions:

$$(x,y_2,y_3,...,y_k)\longmapsto \mathbf{y}'=(y_1',y_2',y_3',...,y_k'),$$

$$y_1'=(y_1-a_1')-x\frac{\sigma_{xy_1}}{\sigma_{xx}},$$

$$y_i'=(y_i-a_i')-x\frac{\sigma_{xy_i}}{\sigma_{xx}},i=2,3,...,k.$$

Now, assuming that $\sigma_{xy_1}=Cov(X,Y_1)\neq 0$, the inverse map

$$\mathbf{y}'=(y_1',y_2',y_3',...,y_k')\longmapsto (x,y_2,y_3,...,y_k)$$

is given by:

$$x=\frac{\sigma_{xx}}{\sigma_{xy_1}}(y_1-y_1'-a_1'),$$

$$y_i=y_i'+a_i'+\frac{\sigma_{xy_i}}{\sigma_{xy_1}}(y_1-y_1'-a_1'),i=2,3,...,k,$$

$$\text{with Jacobian: } |\det\left(\frac{\partial(x,y_2,y_3,...,y_k)}{\partial(y_1',y_2',y_3',...,y_k')}\right)|=\frac{\sigma_{xx}}{\sigma_{xy_1}}.$$

The integrand becomes:

$$D\exp\{-\frac{1}{2}\mathbf{y}'^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\mathbf{y}'\}\left\{\frac{g\left(\frac{\sigma_{xx}}{\sigma_{xy_1}}(y_1-y_1'-a_1')\right)}{g(y_1)}\right\}\frac{\sigma_{xx}}{\sigma_{xy_1}}$$

By assumption,

$$\frac{g\left(\frac{\sigma_{xx}}{\sigma_{xy_1}}(y_1-y_1'-a_1')\right)}{g(y_1)}\leq h(y_1)\text{ for all } y_1$$

for some non-negative function $h(\cdot)$ such that $Eh(Z)<\infty$ when Z has a Gaussian distribution. We therefore have, by dominated convergence theorem

$$\lim_{y_1\to\infty}\int D\exp\{-\frac{1}{2}\mathbf{y}'^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\mathbf{y}'\}\left\{\frac{g\left(\frac{\sigma_{xx}}{\sigma_{xy_1}}(y_1-y_1'-a_1')\right)}{g(y_1)}\right\}\frac{\sigma_{xx}}{\sigma_{xy_1}}d\mathbf{y}'$$

$$=\int D\exp\{-\frac{1}{2}\mathbf{y}'^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\mathbf{y}'\}\lim_{y_1\to\infty}\left\{\frac{g\left(\frac{\sigma_{xx}}{\sigma_{xy_1}}(y_1-y_1'-a_1')\right)}{g(y_1)}\right\}\frac{\sigma_{xx}}{\sigma_{xy_1}}d\mathbf{y}'$$

$$=\int D\exp\{-\frac{1}{2}\mathbf{y}'^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\mathbf{y}'\}\lim_{y_1\to\infty}\left\{\frac{g\left(\frac{\sigma_{xx}}{\sigma_{xy_1}}(y_1-y_1'-a_1')\right)}{g(y_1-y_1'-a_1')}\right\}\lim_{y_1\to\infty}\left\{\frac{g(y_1-y_1'-a_1')}{g(y_1)}\right\}\frac{\sigma_{xx}}{\sigma_{xy_1}}d\mathbf{y}'$$

which, by our assumption on g, in turn equals

$$=\int D\exp\{-\frac{1}{2}\mathbf{y}'^t\mathbf{\Sigma}_{\mathbf{y}|x}^{-1}\mathbf{y}'\}\times\left(\frac{\sigma_{xy_1}}{\sigma_{xx}}\right)^{\alpha}\times\times 1\times\frac{\sigma_{xx}}{\sigma_{xy_1}}d\mathbf{y}'=\left(\frac{\sigma_{xy_1}}{\sigma_{xx}}\right)^{\alpha-1}.\ \square$$

References

[1] S. Abe. Axioms and uniqueness theorem for Tsallis entropy. *Physics Letters A*, 2000.

[2] M. Avellaneda. Minimum entropy calibration of asset-pricing models. *International Journal of Theoretical and Applied Finance.*, 1:447–472, 1998.

[3] M. Avellaneda, C. Friedman, R. Holmes, and D. Samperi. Calibrating volatility surfaces via relative entropy minimization. *Applied Mathematical Finance.*, 4(1):37–64, March 1997.

[4] F. Black and R. Litterman. Asset allocation: combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research*, 1990.

[5] P. Buchen and M. Kelly. The maximum entropy distribution of an asset inferred from option prices. *The Journal of Financial and Quantitative Analysis.*, 31(1):143–159, March 1996.

[6] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, Wiley series in Telecommunications, 1999.

[7] I. Csiszar. A class of measure of informitivity of observation channels. *Periodica Mathematica Hungerica.*, 2(1-4):191–213, 1972.

[8] I. Csiszar. Axiomatic characterization of information measures. *Entropy.*, 10:261–273, 2008.

[9] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer, Application of mathematics-38, 1998.

[10] R. J. V. dos Santos. Generalization of Shannons theorem for Tsallis entropy. *Jounal of Mathematical Physics*, 21, 1997.

[11] P. Dupuis and R. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations.* Wiley, Wiley series in probability and statistics, 1986.

[12] W. Feller. *An Introduction to Probability Theory and its Applications,Vol-2.* John Wiley and Sons Inc., New York, 1971.

[13] J. Gibbs. *Elementary Principles in Statistical Mechanics.* New York: Scribner's, 1902, Reprint-Ox Bow Press, 1981.

[14] P. Glasserman and B. Yu. Large sample prerties of weighted monte carlo estimators. *Operations Research.*, 53(2):298–312, March-April 2005.

[15] L. Golshani, E. Pasha, and Y. Gholamhossein. Some properties of Renyi entropy and Renyi entropy rate. *Information Sciences*, 170:2426–2433, 2009.

[16] E. Jaynes. Information theory and statistical mechanics. *Physics Reviews.*, 106:620–630, 1957.

[17] E. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[18] P. Jizba and T. Arimitsu. The world according to Renyi: Thermodynamics of multifractal systems. *Annals of Physics*, 312:17–59, 2004.

[19] J. Kapur. *Maximum Entropy Models in Science and Engineering.* New Age International Publishers, Wiley series in Telecommunications, 2009.

[20] A. I. Khinchin. *Mathematical foundation of information theory.* Dover, New York, 1957.

[21] A. Meucci. Beyond Black-Litterman in practice. *Risk*, 19(9):114–119, 2006.

[22] A. Meucci. Beyond Black-Litterman: Views on non-normal market. *Risk*, 19(2):96–102, 2006.

[23] A. Meucci. Fully flexible views: theory and practice. *Risk*, 21(10):97–102, 2008.

[24] A. Meucci. Enhancing the Black-Litterman and related approaches:views and stress-test on risk factors. *Journal of Asset Management*, 10(2):89–96, 2009.

[25] J. Mina and J. Xiao. Return to riskmetrics: the evolution of a standard. *RiskMatrics publications*, 2001.

[26] J. Pazier. Global portfolio optimization revisited: a least discrimination alternative to Black-Litterman. *ICMA Centre Discussion Papers in Finance*, July 2007.

[27] E. Qian and S. Gorman. Conditional distribution in portfolio theory. *Financial Analyst Journal.*, 57(2):44–51, March-April 1993.

[28] A. Renyi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960.

[29] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.

[30] C. Tsallis. Possible generalization of Boltzman Gibbs statistics. *Journal of Statistical Physics*, 52:479, 1988.