The Student-t latent factor mixture model with an application to global asset returns

Gideon Magnus*

Senior Quantitative Analyst
Morningstar, Inc.

Preliminary — not to be quoted

May 31, 2016

*I am grateful to Lee Davidson for helpful comments.

Corresponding author:
Gideon Magnus
Morningstar, Inc.
22 W Washington Street
Chicago, IL 60602
USA
Phone: 312-696-6454
e-mail: gideon.magnus@morningstar.com

Abstract We present a method for estimating a model that features both a multivariate mixture of Student-t distributions as well as a latent factor structure. We then apply this method and estimate the joint distribution of excess returns for a large collection of assets spanning the global investable universe. We examine what our estimated model tells us about risk premia in the global asset market, focusing in particular on US stocks. Our results confirm risk premia for momentum, low-beta, and profitability, but not for value, size, and investment. Except for momentum, however, these premia only appear to be exploitable if short positions are allowed.

Keywords Mixture of Student-t distributions · Global portfolio construction · Asset returns · Factor models · Risk premia · Shrinkage

JEL Classification C11 · G10 · G11 · G15 · G17

1 Introduction

This paper has two goals, one theoretical, one empirical. First, we develop a Student-t latent factor mixture model. This model has many potential applications, but we developed it in order to estimate excess asset returns and study risk premia. The passage from theory to practice is non-trivial and we shall deal with several challenges as we progress. In this introduction we provide some background, both on mixture models and asset returns.

Mixture distributions provide a convenient and intuitively appealing way of capturing a wide array of distributional forms, see e.g. Titterington et al. (1985), McLachlan and Basford (1988), McLachlan and Peel (2000), Boldea and Magnus (2009), and Rossi (2014). A prominent class is the mixture of normals. Despite their flexility, however, normal mixtures sometimes fail to capture certain features of the data, in particular behavior in the tails.

A natural extension is the mixture of Student-t distributions. The t-distribution is similar to the normal, but has fatter tails (higher kurtosis). The degree of kurtosis is controlled by a single parameter, the degrees of freedom, and in the limiting case (with infinite degrees of freedom) the t and normal distributions coincide. A t-mixture thus encompasses the normal mixture as a special case. Mixtures of t-distributions have been studied by various authors, including McLachlan and Peel (1998), Peel and McLachlan (2000), Wang et al. (2004), Mclachlan et al. (2006), Gerogiannis et al. (2009), Galimberti and Soffritti (2014), and Zeller et al. (2015).

The model we present here has two features that set it apart from the existing literature. First, we employ a latent factor structure. Conceptually we thus think of the vector of observables as being generated by a (much) smaller collection of factors. Latent factor models provide both computational and conceptual benefits. A high-dimensional model can be estimated quite easily if the number of factors is relatively small. In addition, as in any statistical exercise, the more structure imposed, the more precise the results will be. In our case, the structure provides us with stronger inference especially on assets for which we only have a limited return history.

The second feature is our use of a prior distribution, tailored to our particular application. We use a prior on both the means and standard deviations of each individual asset. This adds plausibility to our results, a crucial feature of any distribution of asset returns; see e.g. Black and Litterman (1992). Without constraints, models of asset returns can quite easily lead to results that are difficult to believe.

Estimation of models with latent factors is typically performed using an Expectation-Maximization (EM) algorithm, which locates the posterior mode. We, in contrast, use direct numerical optimization, where we employ the derivative of the posterior kernel, which we calculate analytically. We also provide a method for simulating the full posterior distribution, as well as a normal approximation to the posterior which appears to be fairly accurate.

In the empirical part we estimate the distribution of global asset class returns and study implied risk premia. A vast literature in empirical finance focuses on identifying the main risk factors that drive excess asset returns; see e.g. Cochrane (2011) for an overview. The typical approach is to select a relatively small number of factor portfolios, and then check whether the returns on a larger collection of assets can be explained by their exposures to these factors. Fama and French (1993), for example, study returns on twenty-five US stock and seven US bond portfolios using a model with three stock and two bond factors. The debate revolves around the most appropriate choice of factors, i.e. the right-hand-side variables in a factor regression.

Risk premia are intrinsically linked to mean-variance optimization: if a factor premium exists, then a mean-variance-optimal investor should tilt his/her portfolio toward assets exposed to that factor. Under heroic theoretical assumptions the mean-variance-optimal (MVO) portfolio coincides with the 'market' portfolio, i.e. the value-weighted total of all tradable assets. A risk premium is defined as a tilt away from the market portfolio and toward a particular exposure.

In this paper we study risk premia using this connection between premia and MVO portfolio tilts. Rather than estimating and comparing results for a potentially large number of factor regressions, we instead estimate the joint distribution of asset returns, from which we then back out the risk premia. We can thus bypass entirely the discussion regarding which factors belong on the right-hand side of a factor regression model.

For our examination of risk premia it is important that we span the global universe of investable assets. If we would omit important assets we could well end up with mistaken inferences. For example, if we only examined US stocks, we might find a premium for (i.e. a tilt toward) small companies. But it could well be that this tilt disappears if we can invest in a global asset portfolio.

A large proportion of the literature on risk premia has focused on US stocks, and we focus on this area of the market as well. Many risk factors have been proposed, analyzed, and debated. We will concentrate on the five most prominent ones: valuation, size, recent performance (i.e. momentum), quality, and volatility/beta. All our US stock data come from the well-known Fama-French database.

There are various ways of quantifying the degree to which a stock price is low, including the book-to-market ratio, earnings-to-price ratio, dividend yield, and cash flow yield. We include five portfolios sorted by book-to-market ratio. For firm size there is less debate about an appropriate definition; we include five portfolios sorted by total market capitalization. References for the value and size premia are numerous, and include Banz (1981), Rosenberg et al. (1985), Chan et al. (1991), and Fama and French (1992, 1993, 1996).

For momentum we include ten portfolios sorted by performance during the preceding year, excluding the most recent month. References for the momentum premium include Jegadeesh and Titman (1993, 2011), Carhart (1997), Fama and French (2012), Novy-Marx (2012), Asness et al. (2013), and Barroso and Santa-Clara (2015).

A recent literature has argued that firm 'quality' carries a risk premium. Quality can be defined in a variety of ways. We include five portfolios sorted by recent investment scaled by firm size. We also include five portfolios sorted by recent operating profitability scaled by the book value of equity. References for the quality premium include Haugen and Baker (1996), Cohen et al. (2002), Fairfield et al. (2003), Titman et al. (2004), Novy-Marx (2013), Hou et al. (2014), Fama and French (2015), and Asness et al. (2015).

Finally, it is thought that there may be a premium for stocks with low market betas, which more or less coincides with low-volatility stocks. We include five portfolios sorted by market beta, where beta is estimated using up to five years of previous monthly returns, with a two-year minimum. References for the low-volatility premium include Ang et al. (2006, 2009), Baker et al. (2011), Frazzini and Pedersen (2014), and Ciliberti et al. (2015).

We also include ten US sector-specific stock portfolios.

Given the scale of the model, estimation of the full posterior is not feasible. We therefore base our results solely on the posterior mode. This requires that the mode provides an adequate approximation to the inference we would obtain if we did incorporate all posterior uncertainty. To check this, we estimate the full posterior, but for a much smaller version of the model. We find posterior uncertainty to be small, and so the estimates based on the mode alone are a close approximation to the posterior predictive distribution.

Our estimation results provide several findings. The MVO portfolio has an annualized expected excess return of 14.42%, standard deviation of 10.49%, and hence a Sharpe ratio of 1.36. It has several sizable short positions, especially in foreign stocks, and a strong net long position in bonds, especially US government bonds and investment grade corporates. There is clear evidence of a momentum premium, a low-beta premium, and a quality premium as measured by profitability. We do not find clear evidence of a value premium or a premium for high-investment stocks. Regarding the size premium, there appears to be a tilt away from the largest firms toward mid-size and smaller mid-size firms.

We also examine MVO portfolios under long-only constraints. Here we find that the constraints are binding for many assets, leading to highly concentrated optimal portfolios. For US stocks the main assets with positive weights are high-momentum stocks and stocks in the utilities and non-durable goods sectors. We estimate various versions of the model and find our results to be fairly robust across alternative specifications.

The plan of this paper is as follows. In Section 2 we present the model and in Section 3 we discuss our estimation method. In Section 4 we introduce our empirical application, while Section 5 contains our main results. In Section 6 we examine the sensitivity of our findings. Section 7 concludes. There are two mathematical appendices.

2 Model and priors

We are interested in an n-dimensional vector of observables x, whose distribution depends on three latent (hence unobserved) variables: a random m-dimensional vector z (the latent factors), a precision parameter v, and a state parameter s. More precisely,

$$x \mid (z, v) \sim \mathrm{N}_n(Bz, (1/v)\Psi), \qquad v \sim \Gamma(\nu/2, \nu/2),$$

and, for $i = 1, \dots, g$,

$$z \mid (v, s = i) \sim \mathrm{N}_m(\mu_i, (1/v)V_i), \qquad \Pr(s = i) = \pi_i.$$

The $n \times m$ matrix B contains the factor loadings associated with the latent factors z.

The latent variables z, v, and s are not considered parameters of our model. Instead, we obtain the distribution of x by integrating them out:

$$p(x) = \sum_{i=1}^{g} \iint p(x \mid z, v)\, p(z \mid v, s = i)\, p(v)\, \Pr(s = i)\, dv\, dz,$$

where the symbol p denotes a density function and will be used generically.

Since we have chosen the parameters in the gamma distribution for v such that 'shape' equals 'rate', the combination of a normal likelihood with a gamma prior on the precision leads to a Student distribution, in this case to the multivariate Student-t mixture distribution (Kotz and Nadarajah, 2004; McLachlan et al., 2006) defined by

$$p(x) = \sum_{i=1}^{g} \pi_i p_i(x) \tag{1}$$

and

$$p_i(x) = \frac{\Gamma[(\nu_i + n)/2]}{\Gamma(\nu_i/2)(\pi\nu_i)^{n/2}} \cdot \frac{|W_i|^{-1/2}}{[1 + \delta_i(x)/\nu_i]^{(\nu_i+n)/2}}, \tag{2}$$

where

$$\delta_i(x) = (x - m_i)' W_i^{-1} (x - m_i) \tag{3}$$

and

$$m_i = B\mu_i, \quad W_i = BV_iB' + \Psi. \tag{4}$$

The π_i are weights satisfying $\pi_i > 0$ and $\sum_i \pi_i = 1$. We shall assume, for simplicity, that Ψ is diagonal and that $\nu_i = \nu > 2$ for all i. We also assume that m, g, and ν are chosen in advance; they are not parameters to be estimated. In practice one would choose m to be much smaller than n, thus reducing the dimension of the model.

The set of parameters is then given by

$$\theta = \{B, \Psi, \{\pi_i, \mu_i, V_i\}_{i=1}^{g}\}.$$

The model as stated is not identified. First, as with any mixture distribution, the labels of the mixture components can be switched arbitrarily. In addition we can, for any non-singular $m \times m$ matrix M, transform the parameters by

$$\mu_i^* = M\mu_i, \qquad V_i^* = MV_iM', \qquad B^* = BM^{-1},$$

and retain the exact same data-generation process. The lack of identification is, however, no concern for purposes of estimation and inference since all parameters that we are interested in are identified.

Letting $\bar{m}$ and $\bar{W}$ denote the mean and variance of x, respectively, we have

$$\bar{m} = \sum_{i=1}^{g} \pi_i m_i, \qquad \bar{W} = \sum_{i=1}^{g} \pi_i \left(\frac{\nu}{\nu - 2} W_i + (m_i - \bar{m})(m_i - \bar{m})' \right). \tag{5}$$

Let $\bar{m}_j$ be the jth component of $\bar{m}$ and let $\bar{w}_j^2$ be the jth diagonal element of $\bar{W}$. We wish to incorporate prior distributions for the means $\bar{m}_j$ and standard deviations (not variances) $\bar{w}_j$ of each x_j. Specifically we assume, for $j = 1, \ldots, n$:

$$\bar{m}_j \sim \mathrm{N}(\rho, \tau^2), \qquad \bar{w}_j \sim \exp(\eta), \tag{6}$$

where $\{\rho, \tau, \eta\}$ are hyperparameters to be chosen in advance. Note that we impose priors not on the parameters themselves, but on functions of the parameters. Note also that the choice of prior for $\bar{w}_j$ implies that the precision $1/\bar{w}_j^2$ follows a Fréchet (or inverse Weibull distribution); see Gumbel (1958). We shall see in Section 6 that our results are not sensitive to the specific shape of the priors. This is why we choose for the simplest among sensible specifications for the prior distributions.

3 Estimation

From here on we write the density (likelihood) $p(x)$ discussed in the previous section as $p(x \mid \theta)$. Given a sample $x_1^{obs}, \ldots, x_T^{obs}$ of independent and identically distributed (iid) observations from this distribution, we write the joint likelihood as $p(X^{obs} \mid \theta)$, where X^{obs} contains all observed data. The log-likelihood of the sample is then $\log p(X^{obs} \mid \theta) = \sum_t \log p(x_t^{obs} \mid \theta)$. In fact, we don't observe all n components of each x_t^{obs}, but only an n_t-dimensional subset $x_{(t)}^{obs} = S_t' x_t^{obs}$, where S_t is an $n \times n_t$ selection matrix. The vectors $x_{(t)}^{obs}$ are now no longer iid, but they are still independent. The log-likelihood thus becomes

$$\log p(X^{obs} \mid \theta) = \sum_{t=1}^{T} \log p(x_{(t)}^{obs} \mid \theta), \tag{7}$$

where it will be convenient, using (1)–(3), to express $p(x_{(t)}^{obs} \mid \theta)$ as

$$p(x_{(t)}^{obs} \mid \theta) = \sum_{i=1}^{g} \pi_i e^{-\lambda_{it}(\theta)/2}, \tag{8}$$

where

$$\lambda_{it}(\theta) = \lambda_{0t} + \log |S_t' W_i S_t| + (\nu + n_t) \log(1 + \delta_{it}/\nu) \tag{9}$$

and

$$\delta_{it} = (x_{(t)}^{obs} - S' m_i)' (S' W_i S)^{-1} (x_{(t)}^{obs} - S' m_i), \tag{10}$$

and λ_{0t} is an (irrelevant) constant.

The log-posterior kernel is then given by the log-likelihood plus the log-prior, which takes the form (ignoring constants):

$$\log p(\theta) = -\frac{1}{2\tau^2} \sum_{j=1}^{n} (\bar{m}_j - \rho)^2 - \eta \sum_{j=1}^{n} \bar{w}_j. \tag{11}$$

Before we turn to estimation, we need to perform three transformations of the parameters. First, the mixture probabilities π_i need to be positive and sum to one. This is achieved by writing

$$\pi_i = \frac{\xi_i^2}{\sum_{j=1}^{g} \xi_j^2} \qquad (i = 1, \ldots, g), \tag{12}$$

where we normalize, without loss of generality, $\xi_1 = 1$. Next, each variance matrix V_i needs to be symmetric and positive definite. We therefore let

$$V_i = \widetilde{V}_i \widetilde{V}_i' + \kappa_1 I_m,$$

where $\widetilde{V}_i$ is lower triangular and κ_1 is a given small positive number. Finally, the diagonal components of Ψ must be strictly positive. We write

$$\Psi = \widetilde{\Psi}^2 + \kappa_2 I_n,$$

where $\widetilde{\Psi}$ is diagonal and κ_2 is a given small positive number. Optimization is performed with respect to $\{B, \widetilde{\Psi}, \{\xi_i\}_{i=2}^{g}, \{\mu_i, \widetilde{V}_i\}_{i=1}^{g}\}$, in total

$$N = n(m+1) + g(m+1)(m+2)/2 - 1$$

parameters. The parameters are ordered in an $N \times 1$ vector θ with components

$$\begin{array}{rl}
\xi_2, \ldots, \xi_g & (g-1 \text{ parameters}) \\
\mu_1, \ldots, \mu_g & (gm \text{ parameters}) \\
\operatorname{vech}(\widetilde{V}_1), \ldots, \operatorname{vech}(\widetilde{V}_g) & (gm(m+1)/2 \text{ parameters}) \\
\operatorname{vec} B & (nm \text{ parameters}) \\
\operatorname{dg}(\widetilde{\Psi}) & (n \text{ parameters}),
\end{array}$$

where, for any square matrix A, $\operatorname{vec} A$ denotes the vector which stacks the columns of A one underneath the other, $\operatorname{vech}(A)$ is obtained from $\operatorname{vec} A$ by deleting all supradiagonal elements of A, and $\operatorname{dg}(A)$ is the vector containing only the diagonal elements of A. In our application, $n = 136$, $m = 30$, and $g = 6$, so that $N = 7191$. The estimation method must therefore be able to deal with a large number of parameters.

Ideally, one would like to generate the complete posterior distribution. This is not feasible, however, given the large dimension of the model. The next best thing is to find key moments of the posterior distribution, in particular the mean $\mu(X^{obs})$ and the variance $\Sigma(X^{obs})$.

Let us denote the score and the Hessian of the log-likelihood as

$$q_{(l)}(\theta, X^{obs}) = \frac{\partial \log p(X^{obs} \mid \theta)}{\partial \theta}, \qquad H_{(l)}(\theta, X^{obs}) = \frac{\partial^2 \log p(X^{obs} \mid \theta)}{\partial \theta \, \partial \theta'},$$

of the log-prior as

$$q_{(p)}(\theta) = \frac{\partial \log p(\theta)}{\partial \theta}, \qquad H_{(p)}(\theta) = \frac{\partial^2 \log p(\theta)}{\partial \theta \, \partial \theta'},$$

and of the log-posterior as

$$q(\theta, X^{obs}) = \frac{\partial \log p(\theta \mid X^{obs})}{\partial \theta}, \qquad H(\theta, X^{obs}) = \frac{\partial^2 \log p(\theta \mid X^{obs})}{\partial \theta \, \partial \theta'}.$$

We estimate the posterior mean by the posterior mode $\bar{\theta}$. The mode is found by maximizing the log-posterior kernel and satisfies the equation

$$q(\bar{\theta}, X^{obs}) = q_{(l)}(\bar{\theta}, X^{obs}) + q_{(p)}(\bar{\theta}) = 0.$$

We use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, as implemented in the `optim` function in `R`, to solve this large non-linear optimization problem. The BFGS algorithm is a quasi-Newton method requiring the gradient but not the Hessian, which is approximated using rank-one updates specified by (approximate) gradient evaluations. We thus need the gradient and we obtain explicit expressions for $q_{(l)}$ and $q_{(p)}$ in Appendices A and B.

We also need good starting values for the parameters. These are obtained as follows. First, we estimate a normal (rather than Student-t) version of the model using only one (rather than six) mixture components, where we initialize the loadings matrix B and factor means μ using an eigendecomposition of the empirical correlation matrix. We then update the values for $\{B, \Psi, \mu\}$ and we use these to calculate 'best guesses' for the latent factors. Next, we estimate a normal multivariate mixture model on these factor guesses with $g = 6$ components, using an Expectation Maximization (EM) algorithm; see e.g. Rossi (2014, p. 10) for details. We run the EM algorithm 25 times, then choose the result with the highest likelihood value. This gives us starting values for all parameters, including $\{\pi_i, \mu_i, V_i\}_{i=1}^{g}$.

To approximate the posterior variance we first note that a second-order Taylor expansion of the log-posterior kernel around $\bar{\theta}$ shows that $\theta \mid X^{obs}$ is approximately normal with mean $\bar{\theta}$ and variance $-H^{-1}(\bar{\theta}, X^{obs})$. Then, we approximate $H(\bar{\theta}, X^{obs})$ as follows:

$$\begin{aligned} H(\bar{\theta}, X^{obs}) &= H_{(l)}(\bar{\theta}, X^{obs}) + H_{(p)}(\bar{\theta}) \approx H_{(l)}(\bar{\theta}, X^{obs}) \\ &\approx \mathrm{E}_{X|\bar{\theta}}\,[H_{(l)}(\bar{\theta}, X)] = -\operatorname{var}_{X|\bar{\theta}}\,[q_{(l)}(\bar{\theta}, X)], \end{aligned}$$

where the first equality is true by definition, the first approximation holds because the prior is approximately linear (in fact approximately constant) for θ close to $\bar{\theta}$, the second approximation is valid if the observed Hessian is well approximated by the expected Hessian, and the final equality follows by second-order regularity (information matrix equality); see Berger (1985, pp. 224–225) for a similar argument and further references.

The variance of the score of the likelihood is not available in closed form, but we can compute it with arbitrary precision by repeatedly sampling from $p(X \mid \bar{\theta})$. The posterior variance $\Sigma(X^{obs})$ is then obtained as the (Moore-Penrose) inverse Q^+ of $Q = \text{var}_{X|\bar{\theta}}\,[q_{(l)}(\bar{\theta}, X)]$. Note that we don't need an explicit expression for the Hessian of the log-likelihood and the log-prior to obtain the posterior variance using this method.

We shall see in Section 6 that this procedure works well in practice and that, in fact, the posterior distribution $p(\theta \mid X^{obs})$ is well approximated by the normal distribution $\phi(\bar{\theta}, Q^+)$.

4 Excess asset returns: data and priors

We shall estimate the joint distribution of excess returns for $n = 136$ assets spanning the global investable universe. We categorize these into nine major classes, listed in Table 1.

	name	index range
1	US stock	1–45
2	Non-US stock	46–96
3	US government bonds	97–114
4	Non-US government bonds	115–117
5	Investment grade bonds	118–123
6	High yield bonds	124–127
7	Inflation-linked bonds	128–129
8	Real Estate Investment Trust	130–131
9	Commodities	132-136

Table 1: Major asset classes

We use monthly returns in excess of the risk-free rate, for which we use 1-month Treasury Bills. The earliest data are from January 1970, the latest from November 2015. We obtain our US stock return data from the Fama and French data library, which can be found at

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

All other data come from Morningstar.

After estimation we apply a modification to the estimated distribution, namely a truncation. Returns cannot be less than -100%, and so we cut off the left tail of each asset's distribution. We also cut off returns that are implausibly high, setting a maximum of 200%. We find that these truncations eliminate 0.4% of the distributional mass. This cut-off is not incorporated in the likelihood function.

Parameter	n	m	g	ν	ρ	τ	η
Value	136	30	6	10	5	4	25

Table 2: Fixed parameters

Table 2 contains our choice of fixed (not to be estimated) parameters, where we have multiplied both ρ and τ by 1200, so that they correspond with annualized percentage scale. Our choice for the number of factors m is based on the eigendecomposition of the empirical correlation matrix: we impose that the first m factors capture 97.5% of the total empirical correlation.

FIGURE 1 HERE

Figure 1 shows the two prior distributions: $\mathrm{N}(5, 16)$ for the prior mean and $\exp(25)$ for the prior standard deviation. We set $\kappa_1 = \kappa_2 = 10^{-5}$.

5 Empirical results

Given the model, priors, and initial conditions, we estimate all $N = 7191$ parameters. We locate the posterior mode through numerical optimization, and base our main results solely on the model implied by the mode. Our primary interest is in the mean $\bar{m}$ and the (diagonal elements of the) variance matrix $\bar{W}$, as defined in (5).

FIGURE 2 HERE

For each of the $n = 136$ assets we plot the mean $\bar{m}_j$ against the standard deviation $\bar{w}_j$ in Figure 2. We also calculate the mean-variance frontier, shown as the solid black line, where we have normalized the risk-free rate at zero. The round black dot indicates the portfolio with the highest Sharpe ratio, which is thus mean-variance optimal, also known as the tangency portfolio. The MVO portfolio has an annualized expected excess return of 14.42%, a standard deviation of 10.49%, and hence a Sharpe ratio of 1.36. The frontier is almost linear, indicating that many alternative portfolios are close to achieving the optimal Sharpe ratio.

Portfolios on the frontier are allowed to have unlimited short positions. Shorting assets is not always straightforward however, and so we also examine the frontier for long-only portfolios. This is shown by the dashed line. Not surprisingly points on the constrained frontier have lower Sharpe ratios, in particular for investors who seek relatively high returns. The highest long-only expected return is provided by Japanese Small Value stocks, namely 10.7% with a Sharpe ratio of 0.26.

Let us consider the points in Figure 2 more closely. Our results are in general agreement with the literature. Bonds tend to have lower returns, especially US government bonds. Corporate bonds have higher returns than US government bonds, and high-yields have higher returns than investment-grades. Stocks, especially foreign stocks, have the highest returns. There is no simple positive relationship between means and standard deviations, and we shouldn't expect there to be, since high returns only compensate for an asset's systematic risk. In particular, some foreign stocks have high volatilities, but not necessarily high expected returns. One stock asset has negative expected returns: the lowest decile of momentum-sorted US stocks. Here the data is simply too strong to convince us otherwise: we have 45 years of data on this asset, with an average annualized excess return of -3.1%.

Estimated means are (almost) all positive, even though the empirical means are sometimes negative. This is because our model induces shrinkage, both through the factor structure and through the priors. A certain amount of shrinkage thus increases the plausibility of our results. Standard deviations are mostly higher than their corresponding sample moments. We also looked at various percentiles, in particular the 5th, 20th, 80th, and 95th percentile. For all percentiles, but especially the 5th and 95th, the model indicates a somewhat greater likelihood of extreme returns than the data. This may, in fact, be a realistic outcome of a model forecasting asset returns.

	total long	total short	net long
US stock	113.7	87.6	26.1
Non-US stock	209.6	219.6	−9.9
US govt bonds	58.1	44.2	13.9
Non-US govt bonds	31.1	14.6	16.5
IG bonds	36.0	4.4	31.7
HY bonds	28.9	0.2	28.7
Infl-linked bonds	0.9	8.2	−7.3
REIT	1.4	0.0	1.4
Commodities	1.3	2.3	−1.0

Table 3: Mean-variance optimal weights (%), unconstrained portfolio

Next, let us study the unconstrained MVO portfolio in more detail. Table 3 shows its composition. The column 'net long' is the difference of the columns 'total long' and 'total short', and adds up to 100%. Portfolio weights are aggregated by major assets. We see that the MVO portfolio has a strong net long position in bonds, especially US government bonds and investment grade corporates.

Expected return	2%	4%	6%	8%
US stock	5.8	12.7	19.9	43.2
Non-US stock	5.2	11.0	17.1	23.6
US govt bonds	68.1	33.5	10.4	0.0
Non-US govt bonds	2.5	5.3	7.3	33.2
IG bonds	10.2	19.4	15.0	0.0
HY bonds	3.8	8.7	15.2	0.0
Infl-linked bonds	3.8	8.4	13.9	0.0
REIT	0.0	0.0	0.0	0.0
Commodities	0.5	1.0	1.2	0.0

Table 4: Mean-variance optimal weights (%), long only

Table 4 shows the compositions of long-only portfolios for four expected returns. When the expected return increases, the portfolio standard deviation increases as well, and as we can see from the dashed line in Figure 2, higher expected returns lead to lower Sharpe ratios. We see from Table 4 that higher expected returns require increasing concentrations in stocks and foreign government bonds.

	unconstrained	2% L/O	4% L/O	6% L/O	8% L/O
size1	2.4	0.0	0.0	0.0	0.0
size2	23.1	0.0	0.0	0.0	0.0
size3	29.1	0.0	0.0	0.0	0.0
size4	4.3	0.0	0.0	0.0	0.0
size5	−63.2	0.0	0.0	0.0	0.0
value1	−15.4	0.0	0.0	0.0	0.0
value2	11.3	0.0	0.0	0.0	0.0
value3	11.4	0.0	0.0	0.0	0.0
value4	15.9	0.0	0.0	0.0	0.0
value5	3.8	0.0	0.0	0.0	0.0
Beta1	35.3	0.0	0.0	0.0	0.0
Beta2	33.2	0.0	0.0	0.0	0.0
Beta3	4.2	0.0	0.0	0.0	0.0
Beta4	−15.3	0.0	0.0	0.0	0.0
Beta5	−21.8	0.0	0.0	0.0	0.0
OP1	−20.6	0.0	0.0	0.0	0.0
OP2	−1.6	0.0	0.0	0.0	0.0
OP3	1.0	0.0	0.0	0.0	0.0
OP4	8.5	0.0	0.0	0.0	0.0
OP5	39.8	0.0	0.0	0.0	0.0
inv1	5.5	0.0	0.0	0.0	0.0
inv2	19.3	0.0	0.0	0.0	0.0
inv3	23.6	0.0	0.0	0.0	0.0
inv4	−1.9	0.0	0.0	0.0	0.0
inv5	−17.7	0.0	0.0	0.0	0.0
Mom1	−23.1	0.0	0.0	0.0	0.0
Mom2	−41.1	0.0	0.0	0.0	0.0
Mom3	−20.3	0.0	0.0	0.0	0.0
Mom4	−12.9	0.0	0.0	0.0	0.0
Mom5	−4.7	0.0	0.0	0.0	0.0
Mom6	5.9	0.0	0.0	0.0	0.0
Mom7	18.2	0.0	0.0	0.0	0.0
Mom8	28.2	0.0	0.0	0.0	0.0
Mom9	30.9	19.8	21.3	21.6	11.9
Mom10	18.1	0.0	0.0	1.1	28.2
NoDur	15.6	53.6	52.1	48.3	20.9
Durbl	0.7	0.0	0.0	0.0	0.0
Manuf	7.1	0.0	0.0	0.0	0.0
Enrgy	3.1	0.0	0.0	0.0	0.0
HiTec	22.0	0.0	0.0	0.0	0.0
Telcm	4.3	0.0	0.0	0.0	0.0
Shops	5.4	0.0	0.0	0.0	0.0
Hlth	−9.7	0.0	0.0	0.0	0.0
Utils	5.3	26.6	26.6	29.0	39.1
Other	−67.1	0.0	0.0	0.0	0.0

Table 5: Mean-variance optimal weights (%). L/O = Long Only

We now focus our attention on the US stock segment of these optimal portfolios. Table 5 shows the optimal portfolio weights, where we have normalized the weights such that in each case they sum to 100%. In this table, x1–x5 refers to portfolios sorted from smallest to largest quintiles for respectively market capitalization, book-to-market ratio, market beta, operating profitability, and investment. Mom1–Mom10 refers to portfolios sorted by momentum deciles. The sectors are respectively Consumer Non-Durables; Consumer Durables; Manufacturing; Oil, Gas, and Coal Extraction and Products; Business Equipment; Telephone and Television Transmission; Wholesale, Retail, and Some Services (Laundries, Repair Shops); Health Care, Medical Equipment, and Drugs; Utilities; Other (Mines, Construction, Building Materials, Transportation, Hotels, Business Services, Entertainment, and Finance).

Do we notice any evidence of risk premia, i.e. tilts? Strictly speaking, in order to distinguish a tilt, we need to know the market portfolio; see e.g. Doeswijk et al. (2014). However, our results provide sufficiently strong indications for tilts, so that a formal comparison with the market does not seem necessary.

The unconstrained optimal portfolio broadly confirms the risk premia proposed in the literature. We see tilts away from large caps and toward mid-size and smaller mid-size firms, except the very smallest firms. Interestingly, the portfolio does not clearly tilt toward high-value stocks. There are, however, clear tilts toward low-beta companies, companies with high profitability, and high momentum. There is no clear tilt toward high investment firms.

The long-only portfolios paint a rather different picture. The constraints are binding for a large number of assets, and the resulting portfolios are highly concentrated. Specifically, the highest momentum stocks are important, especially for investors seeking high expected returns. Besides this, long-only investors should concentrate on stocks in the non-durables and utilities sectors.

6 Sensitivity analysis

In this study we have made many assumptions, and we now ask how sensitive our results are to these assumptions, including our choice of (prior) parameters, our choice of model, and more. We begin by estimating variants of the baseline model, where each variant uses an alternative choice for one of the parameters.

Parameter	n	m	g	ν	τ	η
Baseline value	136	30	6	10	4	25
Alternative value	10	20	3	100	1	50

Table 6: Sensitivity analysis: alternative parameters

Table 6 shows the alternative parameters compared to the baseline parameters, where τ is multiplied by 1200 (as before).

Relevance of Student-t mixtures

We introduced Student-t mixtures, which are more complicated than normal mixtures. Is this added complication worth it? Let us call the model with $\nu = 100$ the 'normal' model. In the normal model, in contrast with our baseline model, the percentiles of the most extreme data points in the sample are (machine) zero. This suggests that, at least for our data, a student mixture is preferable to a normal mixture, since a normal mixture fails to capture suitable behavior in the tails. Hence, the results are sensitive with respect to a simplification to normality.

Relevance of priors

We have introduced priors on the posterior means and standard deviations. How relevant are these priors?

FIGURES 3 AND 4 HERE

If we estimate the model without any priors then some of the means become very large. The same is true (but to a lesser extent) for the standard deviations. This is shown in Figures 3 and 4, respectively. In the model with a tight prior on the means ($\tau = 1$), the means are of course more tightly concentrated around the prior mean (= 5). In the model with a tighter prior on the standard deviations ($\eta = 50$) we find smaller standard deviations, but not too pronounced (obviously an even tighter prior would have a stronger effect). Interestingly, the normal mixture model has the lowest estimated standard deviations, even compared to the model with a stronger prior on standard deviations.

We also experimented with different functional forms for the priors. It turns out that the parameter values are much more important than the functional form.

Number of mixture components

Reducing the number of mixture components from $g = 6$ to $g = 3$ creates a model with fewer parameters, and therefore stronger shrinkage. The effects on the means and standard deviations is not large, however. Hence, there is little indication that we should increase the number of mixture components.

Number of factors

The same is true when we reduce the number of latent factors from $m = 30$ to $m = 20$, as we can see from Figures 3 and 4.

Risk premia

How sensitive are our results regarding the unconstrained MVO portfolio and the implied risk premia?

FIGURE 5 HERE

The optimal weights for all models are shown in Figure 5. There are some differences between the models, although it should be noted that portfolios with different weights could in fact have very similar risk characteristics. If we look at risk premia we find that the normal model, the model without a prior, and the $\eta = 50$ give fairly similar results to the baseline. In contrast, the $g = 3$, $m = 20$, and $\tau = 1$ models appear to be rather different than the baseline model. Interestingly, we see that reducing the number of mixtures g or the number of factors m does not seem to have a large effect on the expected means and standard deviations, but they do affect the risk premia.

MVO portfolio

Suppose our baseline model is 'true', then how bad would investment advice be if we believed a 'wrong' model?

FIGURE 6 HERE

In figure 6 we show the MVO portfolios implied by each model, assuming that our baseline model is the true model. Although using a 'wrong' model may not be a disaster, the differences are far from trivial. The degree to which using the wrong model reduces the optimal Sharpe ratio corresponds to our sensitivity analysis presented above. The 'best wrong' model is the normal model, followed by the model without a prior, and then the model with a stronger prior on standard deviations. The 'worst wrong' model is the model with a stronger prior on means, followed by the model with fewer mixture components, and then the model with fewer latent factors.

Number of assets

We estimate a small-scale version of our model, with only 10 (rather than 136) assets. The main purpose of this exercise is to examine how much posterior uncertainty there is, and to what degree this uncertainty matters for inference. If the posterior is very tight, then the posterior predictive distribution (i.e. the distribution accounting for parameter uncertainty) will be very similar to the distribution implied by the posterior mode alone.

1	US stock: size1
2	US stock: Beta5
3	US stock: Mom4
4	US stock: Health
5	Foreign stock: MSCI Europe Large Value
6	Foreign stock: MSCI World/Financials
7	Foreign stock: S&P Japan LargeMid
8	US government bond: Barclays Agency 20+ Yr
9	US government bond: Citi Treasury Bill 3 Mon
10	High yield bond: Barclays US HY Caa Long

Table 7: Small-scale model: included assets

We select the $n = 10$ assets listed in Table 7, but other selections lead to the same conclusions. We use the same prior parameters as in the full model, with one exception: for the number of latent factors we choose $m = 6$. The heuristic for using this value is the same as for the big model, that is, we impose that the first m factors capture 97.5% of the total empirical correlation.

Following the approximation discussed in Section 3, we approximate the posterior variance as

$$\Sigma(X^{obs}) = Q^{+} + 10^{-12} I, \qquad Q = \mathrm{var}_{X|\bar{\theta}}[q_{(l)}(\bar{\theta}, X)].$$

We use a scaled version of this variance matrix for a Random Walk Metropolis chain whose stationary distribution is the posterior. Specifically, proposal parameter draws are generated from

$$\theta^{new} \sim \mathrm{N}\left(\theta^{old}, (1/10) \times \Sigma(X^{obs})\right).$$

We find that inference based solely on the posterior mode provides an accurate representation of inference based on the full posterior distribution. For instance, asset means, variances, and correlations are virtually identical under the two approaches. Hence, ignoring parameter uncertainty in this paper does not seem to be a great sin.

Finally, we find that the posterior distribution can be closely approximated by a normal approximation, with the posterior mode $\bar{\theta}$ as its mean, and the $\Sigma(X^{obs})$ described above as its variance.

7 Conclusions

In this paper we presented a method for estimating a multivariate mixture of Student-t distributions featuring a latent factor structure. We then applied this method, and estimated the joint distribution of excess returns for a set of assets spanning the global investable universe. We included prior distributions on each asset's mean and standard deviation, which helped to make our results more plausible. Our model sheds new light on the question of risk premia, since the existence of a premium is equivalent to a tilt in a mean-variance optimal portfolio.

Our findings suggest that, at least for US stocks, there are premia for momentum stocks, low-beta stocks, and stocks for companies with high profitability. There is a tilt away from large-cap firms, but not much of a tilt toward the smallest firms. Also, there is not much evidence of a tilt toward value stocks or companies with high levels of investment.

There are a variety of ways in which our model could be expanded in future work. Of particular interest would be to incorporate predictable time-variation in the distribution of returns. Our model is entirely unconditional: the distribution of returns is not affected by any observable variables. This does not mean that it is 'misspecified'; however, it is (to a degree) suboptimal from a forecasting perspective, since it will fail to adapt to information that changes the distribution of returns. The distribution of returns arguably does change, in particular, it is commonly believed that both expected (mean) returns as well as their volatility are to some degree predictable.

Appendix A: Derivative of the log-likelihood

Letting

$$L_t(\theta) = p(x_{(t)}^{obs} \mid \theta) = \sum_{i=1}^{g} \pi_i e^{-\lambda_{it}(\theta)/2}, \tag{13}$$

the log-likelihood is given by $\sum_t \log L_t(\theta)$, using (7) and (8).

It will prove useful to introduce the weights

$$\bar{\pi}_{it} = \frac{\pi_i e^{-\lambda_{it}(\theta)/2}}{\sum_{j=1}^{g} \pi_j e^{-\lambda_{jt}(\theta)/2}}.$$

While π_i denotes the prior probability that the data were generated by mixture component i, $\bar{\pi}_{it}$ can be interpreted as the corresponding posterior probability at time t. We obtain from (13):

$$d \log L_t = \frac{dL_t}{L_t} = \sum_{i=1}^{g} \bar{\pi}_{it} \left(\frac{d\pi_i}{\pi_i} - \frac{d\lambda_{it}}{2} \right), \tag{14}$$

and we thus need to find the differentials of π_i and λ_{it}. We let $\widetilde{\psi} = \mathrm{dg}(\widetilde{\Psi})$, so that $\widetilde{\psi}$ contains the n diagonal components of $\widetilde{\Psi}$. Further, e_i denotes the vector all whose components are zero except the ith which is one. We let $\xi_* = (\xi_2, \ldots, \xi_g)'$ and $S_* = (0 : I_{g-1})$, and we recall that $\xi_1 = 1$. Then $\xi = e_1 + S_*' \xi_*$. Finally, we let

$$p_{it} = (1/\phi_{it}) S_t (S_t' W_i S_t)^{-1} (x_{(t)} - S_t' m_i)$$

and

$$P_{it} = S_t (S_t' W_i S_t)^{-1} S_t' - \phi_{it} p_{it} p_{it}^{'},$$

where $\phi_{it} = (\nu + \delta_{it})/(\nu + n_t)$ and δ_{it} is defined in (10). We then obtain the score as follows.

Proposition 1: The derivative of the log-likelihood (the score) is given by $q_{(l)} = \sum_t q_t$, where q_t is an $N \times 1$ vector with components

$$\frac{\partial \log L_t}{\partial \xi_*} = \sum_{i=1}^{g} (2\bar{\pi}_{it}/\xi_i^2) S_* \left(\xi_i e_i - \pi_i \xi \right),$$

$$\frac{\partial \log L_t}{\partial \mu_i} = B' p_{it}^*, \qquad \frac{\partial \log L_t}{\partial \, \mathrm{vech}(\widetilde{V}_i)} = -\, \mathrm{vech}(B' P_{it}^* B \widetilde{V}_i),$$

$$\frac{\partial \log L_t}{\partial \, \mathrm{vec}\, B} = -\sum_{i=1}^{g} \mathrm{vec} \left(P_{it}^* B V_i - p_{it}^* \mu_i' \right), \qquad \frac{\partial \log L_t}{\partial \widetilde{\psi}} = -\sum_{i=1}^{g} \mathrm{dg}(\widetilde{\Psi} P_{it}^*),$$

where $p_{it}^* = \bar{\pi}_{it} p_{it}$ and $P_{it}^* = \bar{\pi}_{it} P_{it}$.

Proof: After somewhat tedious but straightforward algebra we find from (12):

$$\frac{d\pi_i}{\pi_i} = \frac{2}{\xi_i^2} \left(\xi_i e_i' - \pi_i \xi' \right) d\xi = (2/\xi_i^2) \left(\xi_i e_i - \pi_i \xi \right)' S_*' d\xi_*. \tag{15}$$

With λ_{it} given in (9), using the definitions of p_{it} and P_{it} above, and applying the matrix differential machinery from Magnus and Neudecker (1988), we obtain

$$d\lambda_{it} = \operatorname{tr}(S_t' W_i S_t)^{-1} S_t'(dW_i) S_t + (1/\phi_{it})(d\delta_{it}) = \operatorname{tr} P_{it} dW_i - 2p_{it}' dm_i.$$

Given (4) we find the differentials of m_i and W_i as

$$dm_i = (dB)\mu_i + B d\mu_i$$

and

$$\begin{aligned} dW_i &= (dB)V_i B' + B(dV_i)B' + BV_i(dB)' + d\Psi \\ &= (dB)V_i B' + B(d\widetilde{V}_i)\widetilde{V}_i' B' + B\widetilde{V}_i(d\widetilde{V}_i)'B' + BV_i(dB)' + 2\widetilde{\Psi}(d\widetilde{\Psi}). \end{aligned}$$

This gives

$$\begin{aligned} d\lambda_{it} &= \operatorname{tr} P_{it} dW_i - 2p_{it}' dm_i \\ &= \operatorname{tr} P_{it}(dB)V_i B' + \operatorname{tr} P_{it} B(d\widetilde{V}_i)\widetilde{V}_i' B' + \operatorname{tr} P_{it} B\widetilde{V}_i(d\widetilde{V}_i)'B' \\ &\quad + \operatorname{tr} P_{it} BV_i(dB)' + 2\operatorname{tr} P_{it}\widetilde{\Psi}(d\widetilde{\Psi}) - 2p_{it}'(dB)\mu_i - 2p_{it}' B(d\mu_i) \\ &= -2p_{it}' B(d\mu_i) + 2\operatorname{tr} \widetilde{V}_i' B' P_{it} B(d\widetilde{V}_i) \\ &\quad + 2\operatorname{tr}\left(V_i B' P_{it} - \mu_i p_{it}'\right)(dB) + 2\operatorname{tr} P_{it}\widetilde{\Psi}(d\widetilde{\Psi}), \end{aligned}$$

and hence

$$\begin{aligned} -d\lambda_{it}/2 &= p_{it}' B(d\mu_i) - \operatorname{tr} \widetilde{V}_i' B' P_{it} B(d\widetilde{V}_i) \\ &\quad - \operatorname{tr}\left(V_i B' P_{it} - \mu_i p_{it}'\right)(dB) - \operatorname{tr} P_{it}\widetilde{\Psi}(d\widetilde{\Psi}) \\ &= p_{it}' B d\mu_i - [\operatorname{vech}(B' P_{it} B\widetilde{V}_i)]' d\operatorname{vech}(\widetilde{V}_i) \\ &\quad - \operatorname{vec}\left(P_{it} BV_i - p_{it}\mu_i'\right)' d\operatorname{vec} B - (\operatorname{dg}(\widetilde{\Psi} P_{it}))' d\widetilde{\psi}. \end{aligned} \tag{16}$$

The result follows by inserting the expressions (15) and (16) in (14). ‖

Notice that λ_{it} and $d\lambda_{it}$ do not depend on μ_j and $\operatorname{vech}(\widetilde{V}_j)$ $(j \neq i)$.

Appendix B: Derivative of the log-prior

The log-prior is given by $\log p(\theta)$ in (11):

$$\log p(\theta) = -\frac{1}{2\tau^2}\sum_{j=1}^{n}(\bar{m}_j - \rho)^2 - \eta\sum_{j=1}^{n}\bar{w}_j,$$

which is a function of the components $\bar{m}_j$ of $\bar{m} = \mathrm{E}\, x$ and the diagonal elements $\bar{w}_j^2$ of $\bar{W} = \mathrm{var}(x)$. Of course, other specifications of the prior distributions of $\bar{m}_j$ and $\bar{w}_j^2$ ($j = 1, \ldots, n$) are possible. To facilitate the computation of alternative specifications we present the derivative of the log-prior in a sequence of three propositions.

Regarding $\bar{m} = \sum_i \pi_i m_i = B \sum_i \pi_i \mu_i$, we have

$$d\bar{m} = (dB)\bar{\mu} + B \sum_{i=1}^{g} \pi_i \mu_i (d\pi_i/\pi_i) + B \sum_{i=1}^{g} \pi_i (d\mu_i),$$

where $\bar{\mu} = \sum_i \pi_i \mu_i$. This leads to the following differential.

Proposition 2: The differential of $\bar{m}_j$ is given by

$$d\bar{m}_j = a_j^{(1)\prime} d\xi_* + \sum_{i=1}^{g} a_{ji}^{(2)\prime} d\mu_i + a_j^{(4)\prime} d \operatorname{vec} B,$$

where

$$a_j^{(1)} = (2/\xi'\xi) \sum_{i=1}^{g} (b_j' \mu_i) S_* \left(\xi_i e_i - \pi_i \xi\right), \qquad a_{ji}^{(2)} = \pi_i b_j, \qquad a_j^{(4)} = \bar{\mu} \otimes e_j,$$

and b_j' denotes the jth row of B.

Proof: We have

$$d\bar{m}_j = e_j' d\bar{m} = e_j'(dB)\bar{\mu} + b_j' \sum_{i=1}^{g} \pi_i \mu_i (d\pi_i/\pi_i) + b_j' \sum_{i=1}^{g} \pi_i (d\mu_i),$$

and the result follows from (15). ‖

Regarding $\bar{W}$ we write

$$\bar{W} = \sum_{i=1}^{g} \pi_i W_i^*, \qquad W_i^* = \frac{\nu}{\nu - 2} W_i + (m_i - \bar{m})(m_i - \bar{m})',$$

and we obtain our next result.

Proposition 3: The differential of $\bar{w}_j^2$ is given by

$$d\bar{w}_j^2 = c_j^{(1)\prime} d\xi_* + \sum_{i=1}^{g} c_{ji}^{(2)\prime} d\mu_i + \sum_{i=1}^{g} c_{ji}^{(3)\prime} d \operatorname{vech}(\widetilde{V}_i) + c_j^{(4)\prime} d \operatorname{vec} B + c_j^{(5)\prime} d\widetilde{\psi},$$

where

$$c_j^{(1)} = (2/\xi'\xi)\sum_{i=1}^{g}(e_j'W_i^* e_j)S_*\left(\xi_i e_i - \pi_i\xi\right),$$

$$c_{ji}^{(2)} = 2\pi_i b_j b_j'(\mu_i - \bar{\mu}), \qquad c_{ji}^{(3)} = \frac{2\nu}{\nu-2}\pi_i \operatorname{vech}(b_j b_j'\widetilde{V}_i),$$

$$c_j^{(4)} = 2\operatorname{vec} e_j b_j'\bar{V}, \qquad c_j^{(5)} = \frac{2\nu}{\nu-2}\widetilde{\psi}_j e_j,$$

and

$$\bar{V} = \sum_{i=1}^{g}\pi_i\left(\frac{\nu}{\nu-2}V_i + (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'\right).$$

Proof: We write

$$\begin{aligned}
d\bar{W} &= \sum_{i=1}^{g}(d\pi_i)W_i^* + \sum_{i=1}^{g}\pi_i(dW_i^*) \\
&= \sum_{i=1}^{g}\pi_i(d\pi_i/\pi_i)W_i^* + B\sum_{i=1}^{g}\pi_i\left[(d\mu_i)(\mu_i - \bar{\mu})' + (\mu_i - \bar{\mu})(d\mu_i)'\right]B' \\
&\quad + (dB)\bar{V}B' + B\bar{V}(dB)' + \frac{2\nu}{\nu-2}\widetilde{\Psi}(d\widetilde{\Psi}) \\
&\quad + \frac{\nu}{\nu-2}\sum_{i=1}^{g}\pi_i B(d\widetilde{V}_i)\widetilde{V}_i'B' + \frac{\nu}{\nu-2}\sum_{i=1}^{g}\pi_i B\widetilde{V}_i(d\widetilde{V}_i)'B'.
\end{aligned}$$

This implies that

$$\begin{aligned}
d\bar{w}_j^2 &= \sum_{i=1}^{g}\pi_i(e_j'W_i^* e_j)a_i'(d\xi_*) + 2\sum_{i=1}^{g}\pi_i(\mu_i - \bar{\mu})'b_j b_j'(d\mu_i) \\
&\quad + \frac{2\nu}{\nu-2}\sum_{i=1}^{g}\pi_i(\operatorname{vech}(b_j b_j'\widetilde{V}_i))'d\operatorname{vech}(\widetilde{V}_i) \\
&\quad + 2(\operatorname{vec} e_j b_j'\bar{V})'(d\operatorname{vec} B) + \frac{2\nu}{\nu-2}\widetilde{\psi}_j e_j'(d\widetilde{\psi}),
\end{aligned}$$

and the result follows. ‖

We now have all ingredients to obtain the derivative of the log-prior.

Proposition 4: The derivative of the log-prior is given by the $N \times 1$ vector

$q_{(p)}$ with components

$$\frac{\partial \log p(\theta)}{\partial \xi_*} = -(1/\tau^2)\sum_j (\bar{m}_j - \rho) a_j^{(1)} - (\eta/2)\sum_j (1/\bar{w}_j) c_j^{(1)},$$

$$\frac{\partial \log p(\theta)}{\partial \mu_i} = -(\eta/2)\sum_j (1/\bar{w}_j) c_{ji}^{(3)},$$

$$\frac{\partial \log p(\theta)}{\partial \operatorname{vech}(\widetilde{V}_i)} = -(1/\tau^2)\sum_j (\bar{m}_j - \rho) a_{ji}^{(2)} - (\eta/2)\sum_j (1/\bar{w}_j) c_{ji}^{(2)},$$

$$\frac{\partial \log p(\theta)}{\partial \operatorname{vec} B} = -(1/\tau^2)\sum_j (\bar{m}_j - \rho) a_j^{(4)} - (\eta/2)\sum_j (1/\bar{w}_j) c_j^{(4)},$$

$$\frac{\partial \log p(\theta)}{\partial \widetilde{\psi}} = -(\eta/2)\sum_j (1/\bar{w}_j) c_j^{(5)}.$$

Proof: The differential is

$$\begin{aligned} d\log p(\theta) &= -(1/\tau^2)\sum_j (\bar{m}_j - \rho)(d\bar{m}_j) - \eta \sum_j (d\bar{w}_j) \\ &= -(1/\tau^2)\sum_j (\bar{m}_j - \rho)(d\bar{m}_j) - (\eta/2)\sum_j (1/\bar{w}_j)(d\bar{w}_j^2), \end{aligned}$$

and the results follow from Propositions 2 and 3. ‖

References

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1), 259–299.

Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. (2009). High idiosyncratic volatility and low returns: International and further U.S. evidence. *Journal of Financial Economics*, 91(1), 1–23.

Asness, C. S., Frazzini, A., and Pedersen, L. H. (2015). Quality minus junk. Working paper, AQR.

Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance*, 68(3), 929–985.

Baker, M., Bradley, B., and Wurgler, J. (2011). Benchmarks as limits to arbitrage: Understanding the low volatility anomaly. *Financial Analysts Journal*, 67(1), 40-54.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, 9(1), 3–18.

Barroso, P. and Santa-Clara, P. (2015). Momentum has its moments. *Journal of Financial Economics*, 116(1), 111–120.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5), 28–43.

Boldea, O. and Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, 104(488), 1539–1549.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, 52(1), 57–82.

Chan, L. K. C., Hamao, Y., and Lakonishok, J. (1991). Fundamentals and stock returns in Japan. *The Journal of Finance*, 46(5), 1739–1764.

Ciliberti, S., Lempérière, Y., Beveratos, A., Simon, G., Laloux, L., Potters, M., and Bouchaud, J. P. (2015). Deconstructing the low-vol anomaly. Capital Fund Management, Paris, France.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4), 1047–1108.

Cohen, R. B., Gompers, P. A., and Vuolteenaho, T. (2002). Who underreacts to cash-flow news? Evidence from trading between individuals and institutions. *Journal of Financial Economics*, 66(2-3), 409–462.

Doeswijk, R., Lam, T., and Swinkels, L. (2014). The global multi-asset market portfolio, 1959–2012. *Financial Analysts Journal*, 70(2), 26–41

Fairfield, P. M, Whisenant, J. S, and T. L Yohn (2003). Accrued earnings and growth: Implications for future profitability and market mispricing. *The Accounting Review*, 78(1), 353–371.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2), 427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33, 3–56.

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, 51(1), 55–84.

Fama, E. F. and French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, 105(3), 457–472.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22.

Frazzini, A. and Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, 111(1), 1–25.

Galimberti, G. and Soffritti, G. (2014). A multivariate linear regression analysis using finite mixtures of t distributions. *Computational Statistics & Data Analysis*, 71, 138–150.

Gerogiannis, D., Nikou, C., and Likas, A. (2009). The mixtures of Student's *t*-distributions as a robust framework for rigid registration. *Image and Vision Computing*, 27(9), 1285–1294.

Gumbel, E. J. (1958). *Statistics of Extremes.* New York: Columbia University Press.

Haugen, R. A. and Baker, N. L. (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics*, 41(3), 401–439.

Hou, K., Xue, C., and Zhang, L. (2014). A comparison of new factor models. *NBER Working Paper* 20682. National Bureau of Economic Research.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65–91.

Jegadeesh, N. and Titman, S. (2011). Momentum. *Annual Review of Financial Economics*, 3(1), 493–509.

Kotz, S. and S. Nadarajah (2004). *Multivariate t Distributions and Their Applications.* New York: Cambridge University Press.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, First revision 1991, Second edition 1999, Third edition (free download) 2007. Chichester/New York: John Wiley and Sons.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York / Basel: Marcel Dekker Inc.

McLachlan, G. J., S.-K. Ng, and R. Bean (2006). Robust cluster analysis via mixture models. *Australian Journal of Statistics*, 35, 157–174.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In: A. Amin, D. Dori, P. Pudil, and H. Freeman (eds.), *Advances in Pattern Recognition*, *Lecture Notes in Computer Science*, Vol. 1451, pp. 658–666. Berlin / Heidelberg: Springer.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley.

Novy-Marx, R. (2012). Is momentum really momentum? *Journal of Financial Economics*, 103(3), 429–453.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, 108(1), 1–28.

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4), 339–348.

Rosenberg, B., Reid, K., and Lanstein, R. (1985). Persuasive evidence of market inefficiency. *Journal of Portfolio Management*, 11(3), 9–17.

Rossi, P. E. (2014). *Bayesian Non- and Semi-Parametric Methods and Applications*. The Econometric and Tinbergen Institutes Lectures. Princeton, NJ: Princeton University Press.

Titman, S., Wei, K. C. J., and Xie, F. (2004). Capital investments and stock returns. *Journal of Financial and Quantitative Analysis*, 39(4), 677–700.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley and Sons.

Wang, H. X., Zhang, Q. B., Luo, B., and Wei, S. (2004). Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters*, 25(6), 701–710.

Zeller, C., Cabral, C., and Lachos, V. (2015). Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *TEST*, 1–22.

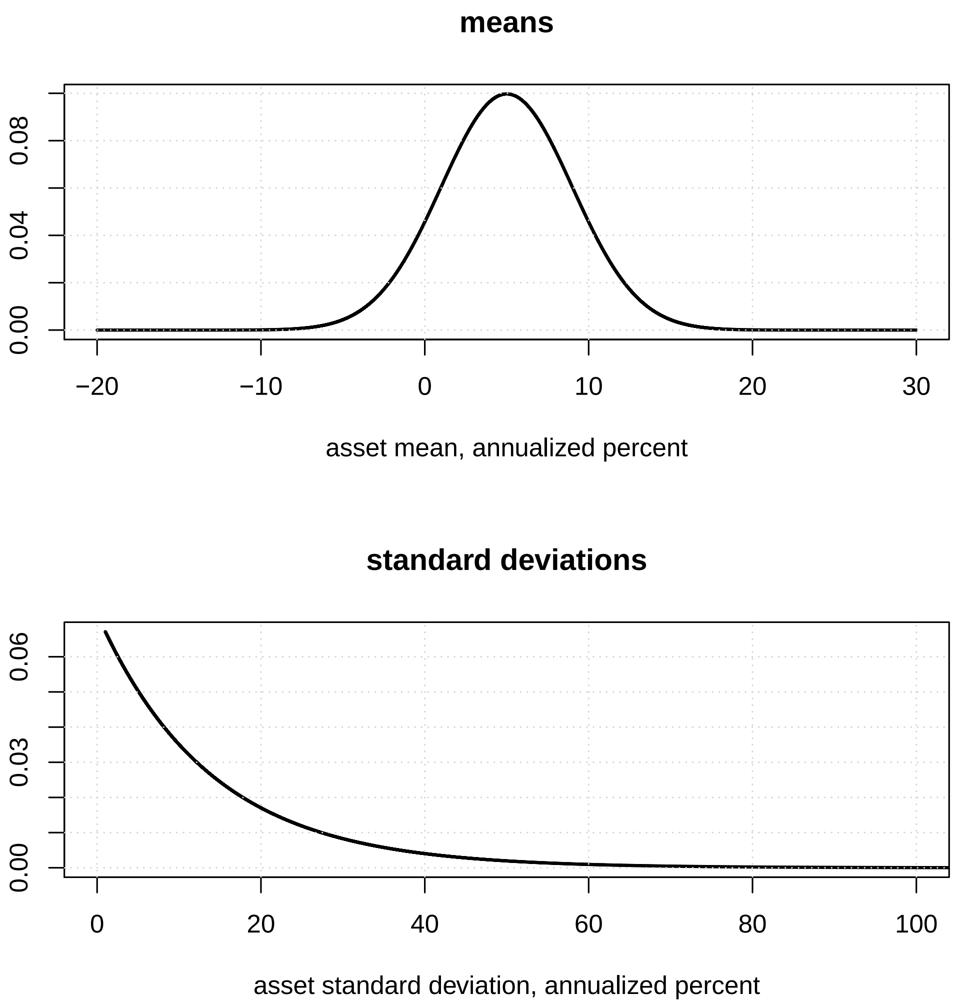

Figure 1: Prior for mean and standard deviation

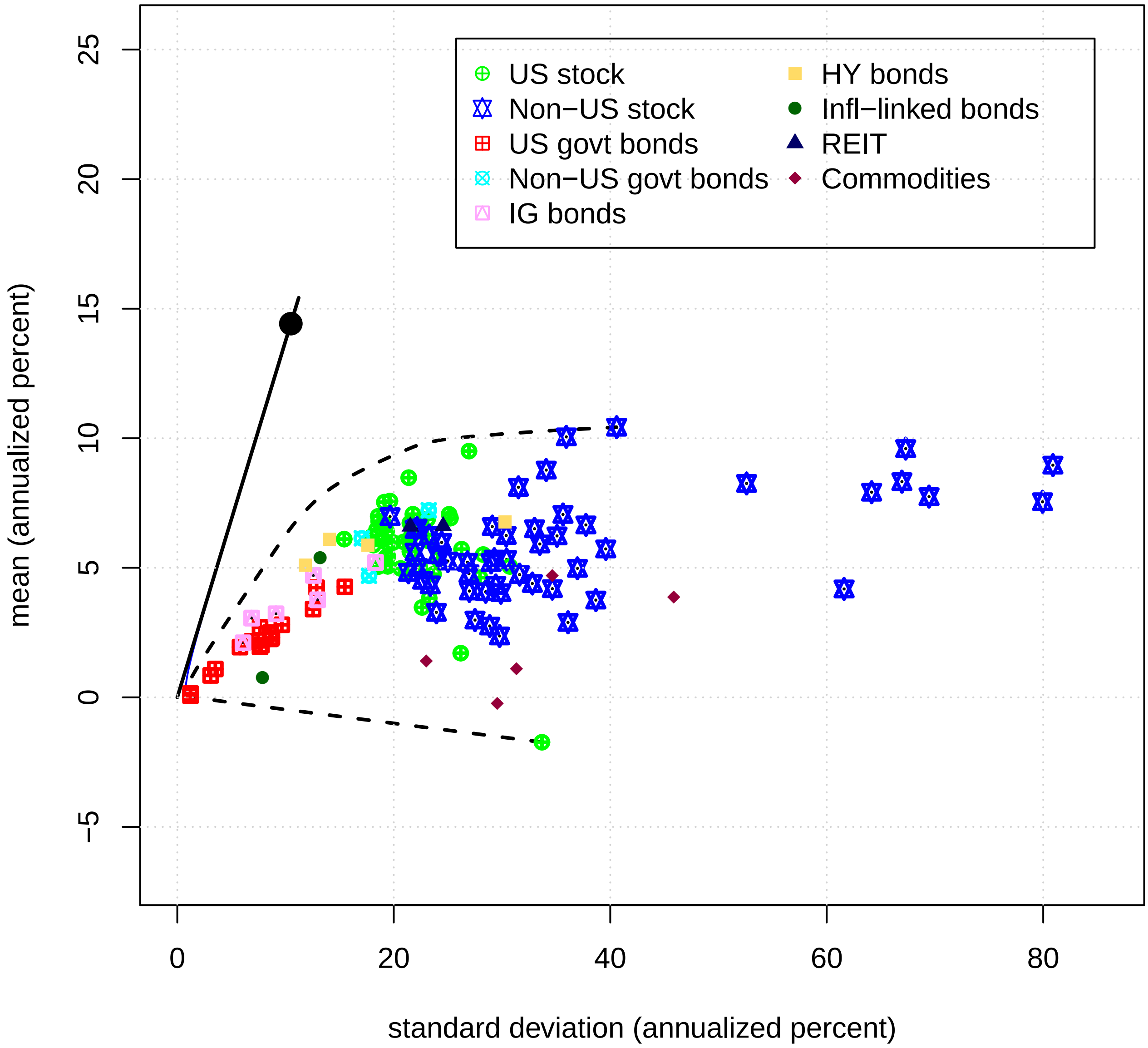

Figure 2: Mean-variance frontier

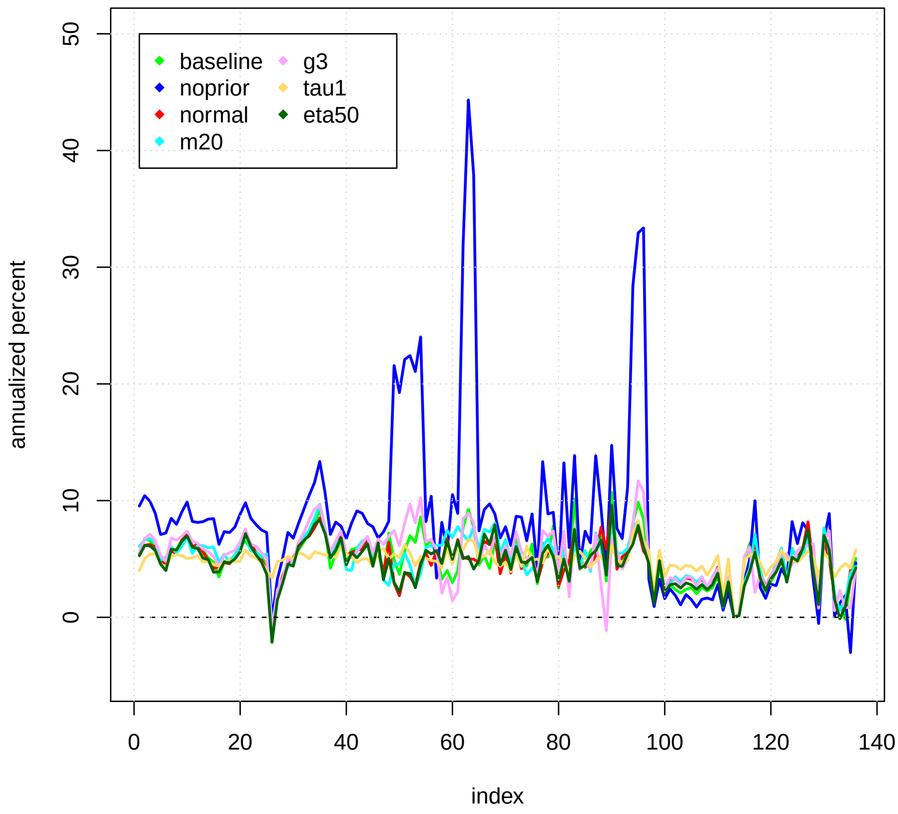

Figure 3: Asset means across models

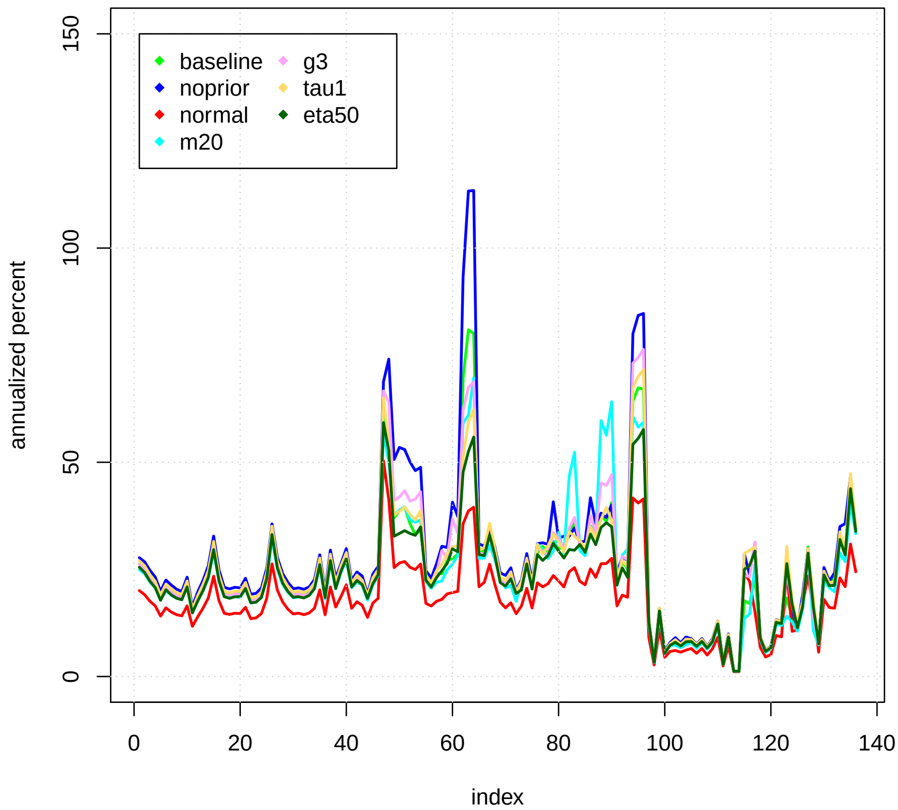

Figure 4: Asset standard deviations across models

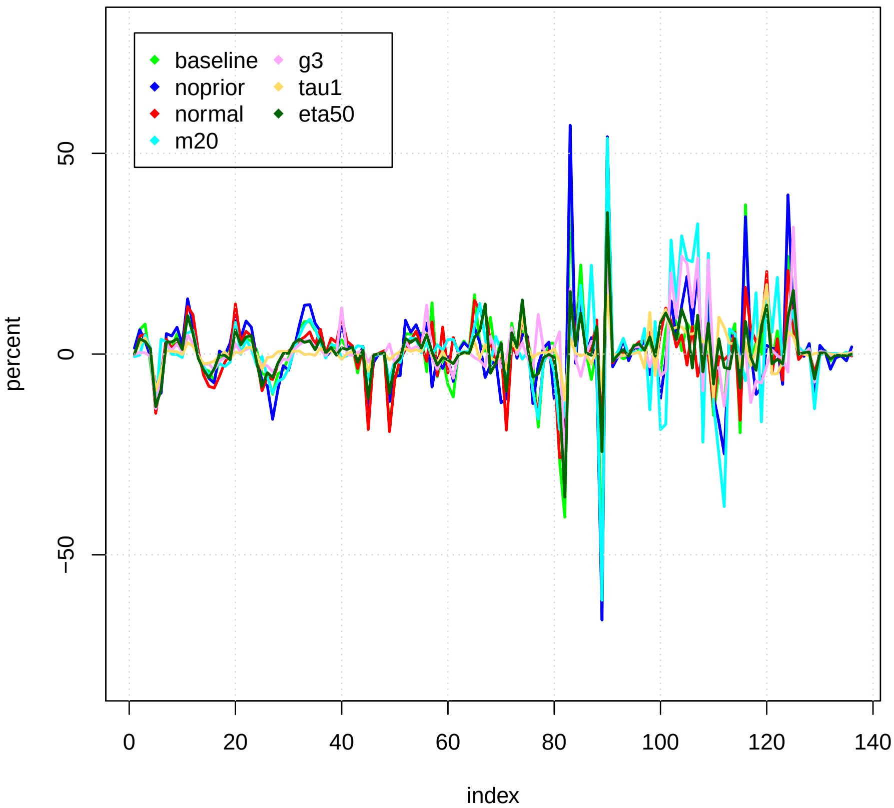

Figure 5: Optimal weights across models (unconstrained portfolios)

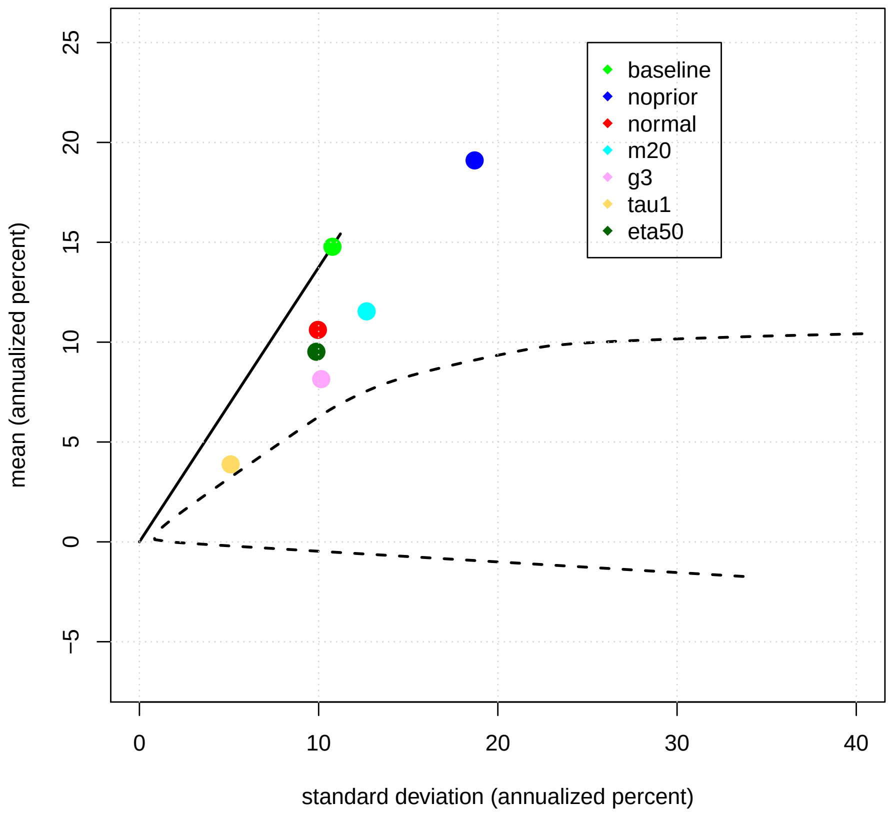

Figure 6: Mean-variance characteristics of portfolios that are mean-variance optimal according to each model