ESSAYS ON DECISION MAKING UNDER UNCERTAINTY: STOCHASTIC DOMINANCE

A Thesis presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

LONGHAO ZHUO

Dr. David Kaplan, Thesis Supervisor

DECEMBER 2017

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ESSAYS ON DECISION MAKING UNDER UNCERTAINTY:
STOCHASTIC DOMINANCE

presented by Longhao Zhuo,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. David Kaplan

Dr. J. Isaac Miller

Dr. Peter Mueser

Dr. Dongchu Sun

ACKNOWLEDGMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my parents and wife.

I would like to express my deepest gratitude to my advisor, Dr. David Kaplan, for his endless patience, enthusiastic guidance, and inspiring discussion. I would also like to thank Dr. Miller, Dr. Mueser, Dr. Ni and Dr. Sun for their invaluable comments and feedback.

I would like to thank many people I met in Columbia Missouri for their friendship, services and feedback. Special thanks go to the late Dr. George Wagner, who encouraged, if not pushed, me all the time for early graduation.

Last but not the least, I would like to thank my parents for their unconditional love and support in the ups and downs of my life, also thank my wife for her smile and assurance, and for making the hard decisions easy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Decision making under uncertainty is certainly the most important task of an economics agent and it is often a very difficult one. In most cases, the goal of further analysis of uncertainty is not necessarily to reduce it, but to better understand it and its implications for the decision makers. In this regard, this dissertation focuses on an useful concept called stochastic dominance (SD) and its econometric inference on various applications. SD is not only a comprehensive measure of risk and uncertainty, but has strong implications for the welfare and utility of economic agents.

The dissertation consists of three chapters. The first chapter proposes a nonparametric Bayesian method for providing probabilistic measurement on stochastic dominance (SD) of any order. We use the approach of Rubin (1981) for implementing the model of Ferguson (1973, 1974) with an improper noninformative Dirichlet process prior. The posterior is not only logically coherent among all orders of SD, but relevant for decision making under uncertainty in welfare analysis. Monte Carlo results show our Bayesian procedure outperforms other nonparametric frequentist tests in terms of Bayes risk in many cases. We extend the model to consider sample weights and clustered sampling error. The results are illustrated using data from the Panel Study of Income Dynamics.

Chapter two discusses the improper use of ordinal data as a measure of health in empirical research. In particular, we focus on a kind of questions, in which its qualitative nature in measurement restricts the scope of questions it can answer. To illustrate this limitation, we present two examples using ordinal self-reported health status (SRHS). In the first example of age effect on health, we find SRHS alone may not be adequate for inferring health inequality or dispersion. It shows only that average health declines with age. We also study the inter-cohort trend in health inequality (i.e., cohort effect). The main findings are 1) the elderly is reported healthier today

than before and 2) the health of the Black elderly is largely improved over years and becomes less unequal since year 2010. Appropriate statistical inferences on ordinal data are recommended.

Chapter three proposes and implements an enhanced indexing strategy based on the stochastic dominance (SD) decision criteria, nonparametric Bayesian (NPB) inference and stochastic optimization algorithm. SD and NPB share a distribution-free assumption framework which allows a robust approach for non-normal return distributions. Further, NPB provides the probabilistic basis for optimization when uncertainty is present in problems of decision making. In particular, SD/NPB method can be applied by constructing an optimization problem constrained by stochastic dominance relations. We discuss the uncertainties around these relation and find the optimal portfolio using the mixed-integer linear programming (MILP) algorithm. Our method yields important ex-ante performance improvements relative to heuristic diversification, Mean-Variance optimization and widely-used Standard&Poor 500 index (SP500). Relative to SP500, our method improves average out-of-sample return by more than nine percentage points per annum, with higher Sharpe ratio, three-month re-balancing and no short sales.

Chapter 1

Nonparametric Bayesian inference on stochastic dominance

1.1 Introduction

We obtain posterior probabilities of first-order and higher-order stochastic dominance relationships by using a nonparametric Bayesian method under an improper noninformative prior. The basic model is due to Ferguson (1973, 1974), Rubin (1981), and Banks (1988). We further extend it to allow sample weights Lo (using 1993) and/or clustered errors. Three valuable features of this particular nonparametric Bayesian approach, as pointed out by Chamberlain and Imbens (2003, p. 12), are maintained here. First, it simultaneously provides probabilistic measurement on all possible dominance relationships. Second, it does not assume any fully parametric specification for the distributions being compared for stochastic dominance. Third, the noninformative prior reflects minimal subjective opinion and makes implementation easy.

Stochastic dominance (SD) has been fundamental in analyzing income inequality and social welfare because it does not impose assumptions on the functional forms of income distributions or specifications of utility function (see, e.g., Ravallion (1994) and Deaton (1997)). The SD approach is appealing since it can provide a robust comparison between distributions, and such ordering holds true for a large set of preferences. For example, the income distribution in society A second-order stochastically dominates the distribution in another society if and only if A has a higher welfare given any non-satiated (non-decreasing) and risk-averse (concave) social utility.

SD is also helpful when studying treatment effects. For example, we may wish to know whether the treated potential outcome distribution first-order or second-order stochastically dominates the untreated distribution. This is much more informative than the average treatment effect and more relevant to the policy question of whether to adopt the treatment. The methods in this paper may be applied not only to data from randomized experiments, but also conditional distributions (under unconfoundedness) and regression discontinuity designs, applying results from Canay and Kamat (2017) as in Goldman and Kaplan (2017, §6).

Most previous methods for SD inference are frequentist. Among others, Davidson and Duclos (2000) characterize the null hypothesis of SD as inequality constraints at a fixed number of arbitrary chosen points and derive the asymptotic sampling distribution of related test statistics. Barrett and Donald (2003) propose a more powerful method by testing the inequalities at all points in the support of the distributions. They use the bootstrap to simulate the critical value of a Kolmogorov–Smirnov type test for SD.

Others in the frequentist literature consider the null hypothesis of *non*dominance (*non*SD). Kaur et al. (1994) originally propose such a null to avoid difficulties in constructing a valid rejection region without simulation/bootstrap. They show that when the (least favorable) null of *non*SD is true, the limiting distribution of their intersection–union t-test statistic is a standard normal distribution. Davidson and Duclos (2013) further advocate the *non*SD null since rejecting it provides stronger evidence of SD than failing to reject a null of SD. In fact, the null of *non*SD is widely used in finance. For example, Post (2003) tests the SD efficiency of a given portfolio by considering whether it is dominated by any other feasible portfolio.

Although the nonparametric Bayesian approach to SD inference has been largely unexplored, there are (at least) three reasons to study it. First, the Bayesian framework deals with non-dominance and dominance simultaneously and provides their posterior probabilities based on the data. It treats *non*SD and SD in a full probability model, in which the posterior probabilities of the three possible dominance relationships (i.e., $X\text{SD}Y$, $Y\text{SD}X$, and *non*SD) sum to one. Second, it remains unclear if Bayesian and frequentist inferences on SD can be reconciled. Kaplan and Zhuo (2017) find that even when the frequentist sampling distribution and Bayesian posterior distribution are asymptotically equivalent, frequentist and Bayesian approaches may reach opposite conclusions on a joint test of multiple inequalities. SD of any order can be written as such a set of inequalities. Third, the reliability of parametric inferences are often overshadowed by misspecification of the likelihood. To increase robustness within the Bayesian framework, Lubrano and Ndoye (2016) and Lander et al. (2016) suggest income be modeled as a finite mixture distribution, which is a collection of simple parametric distributions. But it is potentially flawed (and vulnerable to manipulation) by the ad hoc number of components and choice of parametric distribution.

In this article, we use a simple nonparametric Bayesian method for inference on the cumulative distribution function (CDF) $F(\cdot)$, which is treated as an infinite-dimensional parameter, for each population of interest. As in Rubin (1981), an improper noninformative Dirichlet process (DP) prior in the framework of Ferguson (1973) is used over $F(\cdot)$ to facilitate the computation of its posterior distribution. In particular, we take the limit as the DP prior's hyperparameter approaches the zero function. Rubin (1981) names this the *Bayesian bootstrap* and discusses sampling from the posterior distribution of $F(\cdot)$ via Monte Carlo simulation. Banks (1988) provides a continuity correction for use with continuous CDFs. Lo (1987) and Weng (1989) show the centered and scaled posterior of $F(\cdot)$ to have the same asymptotic limit regardless of the DP prior used, so the effect of the improper DP prior (versus using another DP prior) is small in large datasets. With independent samples, draws from each posterior may be taken independently, and the posterior probability of any SD relationship is the proportion of draws in which it holds.

Although motivated by Chamberlain and Imbens (2003), who discuss the merits of the Bayesian bootstrap in economics, our approach is more than a simple extension of theirs. Chamberlain and Imbens (2003) argue that the posterior distribution is useful in accommodating parameter uncertainty for decision making through an instrumental variable example and that the posterior provides better inference when the asymptotic approximation to the sampling distribution is poor, through a quantile regression example. Three differences in our paper should be noted. First, the CDF itself is of interest here, whereas it is only a nuisance parameter in their study. In particular, since SD relations depend on the tails, the continuity correction for continuous CDFs is very important, since otherwise SD may be falsely rejected with 100% probability at the sample maximum or minimum in some cases. Second, the notion of SD involves inequalities at infinitely many points. We find our nonparametric Bayesian conclusions are strikingly different from the frequentist ones on these joint inequality constraints. Such findings are also seen in Kline (2011) and Kaplan and Zhuo (2017). Third, we develop extensions to accommodate sampling weights and clustering in survey data.

Two specific problems concerning empirical SD testing are unequal probability sampling and a clustering structure. First, ignoring sample weights in survey data can lead to incorrect inferences about the population of interest. To incorporate sample weights into the Bayesian bootstrap, Lo (1993) introduces the normalized weighted gamma process prior. Under this prior, he shows the modified Bayesian bootstrap approximation to the posterior distribution of the mean is accurate when probabilities of selection, the inverse of sample weights, are modeled through the weighting distribution model (Rao, 1965). It is not required that one rescales weights inside each bootstrap sample, which is inevitable in frequentist bootstraps (Rao and Wu, 1988). Second, clustering is about within-group/cluster correlation and hence is a common phenomenon in economics, especially with panel data. But such correlation violates the iid assumption in basic bootstrap techniques. Cameron et al. (2008) consider the cluster bootstrap that resamples the whole cluster as one sampling unit. Similar to the cluster bootstrap, the Bayesian bootstrap can also simulate at the level of clusters, instead of individual observations.

Section 1.2 contains the algorithms of implementing our nonparametric Bayesian analysis in the basic setting as well as with complex sampling. Section 1.3 contains a brief review of SD and formulation of the corresponding hypotheses. Sections 1.4 and 1.5 contain simulation results and empirical examples, respectively. Acronyms and abbreviations used include those for Bayesian bootstrap (BB), cumulative distribution function (CDF), Dirichlet process (DP), random variable (RV) and jth-order stochastic dominance (SDj). Notationally, random and non-random vectors are respectively typeset as, e.g., $\mathbf{X}$ and $\mathbf{x}$, while random and non-random scalars are typeset as X and x. The Dirichlet distribution with parameters $a_1, \ldots, a_K$ is written $\mathrm{Dir}(a_1, \ldots, a_K)$.

1.2 Bayesian bootstrap method

We present a nonparametric Bayesian method where a noninformative prior is used to derive the posterior of a CDF, which serves as a crucial function in SD inference. It is "nonparametric" because the true CDF $F(\cdot)$ is an infinite-dimensional parameter.

For such parameter, the most commonly used prior is the Dirichlet process prior proposed by Ferguson (1973). In particular, Rubin (1981) coins the term “Bayesian bootstrap” (BB) when using an improper and noninformative prior, developing a Bayesian analogy to the frequentist bootstrap. In fact, the BB is a valid frequentist bootstrap, but we use it with a Bayesian interpretation.

In section 1.2.1, the basic method of the BB with iid data is introduced. We then discuss the similarities between Bayesian and frequentist uniform confidence/credible bands for the CDF in section 1.2.2. In sections 1.2.3 and 1.2.4, BB is further extended to account for sample weights and clustering, respectively.

1.2.1 Basic method

Let $\{X_i\}_{i=1}^n \overset{iid}{\sim} F$. To learn about the unknown $F \in \mathcal{F}$ from a Bayesian perspective, a prior is specified over the set $\mathcal{F}$ of possible CDFs, and then the posterior distribution of F on $\mathcal{F}$ is computed given the sample $\{X_i\}_{i=1}^n$.

A preliminary example is to address the posterior distribution of discrete F having the finite support $\{d_k\}_{k=1}^K$. The prior of F is often assumed to have a Dirichlet distribution, which is a multivariate distribution having support over a $K \times 1$ vector whose entries are real numbers in the interval $[0, 1]$ and sum together to 1. Given the data sampled from F, the posterior distribution is obtained as usual, made simple by the conjugacy of the Dirichlet distribution prior and the multinomial likelihood.

When it comes to continuous F, the prior distribution is of great importance, since the CDF is an infinite-dimensional parameter. So, the Dirichlet distribution prior fails to manage the task since it can only account for a fixed number of parameters. One solution is to use the Dirichlet process prior, defined by Ferguson (1973, 1974) for the general nonparametric Bayesian framework. The Dirichlet process (DP) is a random probability measure, which defines the joint distribution for every measurable partition $(\mathcal{B}_1, \ldots, \mathcal{B}_K)$ on the sample space. That is, DP defines the distribution of

$\big(\mathrm{P}(\mathcal{B}_1), \ldots, \mathrm{P}(\mathcal{B}_K)\big)$ for all K. For a particular K and the corresponding partition, the marginal distribution is Dirichlet. Roughly speaking, DP is a joint distribution over all these (varying K) marginalized distributions. It is uniquely controlled by the parameter $\alpha(\cdot)$. Ferguson (1973, p. 217) shows if the prior knowledge of F can be summarized by a DP with hyperparameter $\alpha(\cdot)$ and the sample $\mathbf{X}$ of size n is drawn from F, then the posterior distribution of F is also a DP with updated parameter $\alpha(\cdot) + \sum_{j=1}^{n} \delta_{x_j}(\cdot)$, where $\delta_x(A) = \mathbb{1}\{x \in A\}$.

The BB in Rubin (1981) is a practical implementation of this nonparametric Bayesian framework. It uses an arguably noninformative DP prior by letting its hyperparameter $\alpha(\cdot) \to 0$.[1] Under this type of peculiar prior, the DP posterior distribution becomes much more tractable: $\alpha(\cdot)$ only updates at observed values in the sample; and, at the rest of values in the support, $\alpha(\cdot)$ are all zero with the posterior probability one. It is a Dirichlet posterior distribution. A great advantage of Rubin's BB is that the posterior distribution of F can be approximated directly by a Monte Carlo simulation from this Dirichlet posterior distribution. Each simulation corresponds to a realization of the discrete distribution for $p_1 = F(X_{n:1})$, $p_2 = F(X_{n:2}) - F(X_{n:1})$, etc., up to $p_n = F(X_{n:n}) - F(X_{n:n-1})$, where $(X_{n:1}, \ldots, X_{n:n})$ are the order statistics of the sample $\mathbf{X}$. Also, the Dirichlet-distributed weights $\mathbf{p} = (p_1, \ldots, p_n)$ in BB lead to smoothing when compared to the original (multinomial weights) bootstrap of Efron (1979).[2]

Asymptotic studies of BB have focused on the (lack of) influence of the prior in large samples. Major contributions are made by Lo (1987) and Weng (1989), among others. Lo (1987) shows, for any DP prior over F, its posterior distribution can be first-order approximated by the conditional distribution of $T_n(\cdot)$ used in Rubin (1981).

[1] For the DP prior, Gelman et al. (2014) claim $\alpha(\cdot)$ reveals in some sense a prior sample size.

[2] Another way to view it is from the perspective of a smoothing bootstrap as in Lancaster (2003): instead of resampling data directly, it assigns a random probability to each observation, and random probabilities are simulated from a posterior distribution.

The random CDF $T_n(\cdot)$ is defined as

$$T_n(x) \equiv \sum_{k=1}^{n} p_k \, \mathbb{1}\{X_{n:k} \le x\}, \tag{1.1}$$

where $p_k \equiv \mathrm{P}(X = X_{n:k})$ as above. The conditional distribution of $T_n(\cdot)$ (given data sample $\mathbf{X}$) is obtained by simulating $(p_1, \ldots, p_n)$ from the posterior distribution many times, say B times, to have $T_n^1, \ldots, T_n^B$.

Lo (1987, Thm. 2.1) proves, for almost all sample sequences $\mathbf{X}$, the recentered and rescaled $T_n(\cdot)$ converges to a Brownian bridge. That is, with $\rightsquigarrow$ denoting weak convergence,

$$\sqrt{n}\big(T_n(\cdot) - \hat{F}(\cdot)\big) \rightsquigarrow \mathfrak{B}\big(F(\cdot)\big) \tag{1.2}$$

where $\hat{F}(\cdot)$ is the empirical CDF from data and F is the true CDF, and $\mathfrak{B}$ is a standard Brownian bridge. This result holds for any DP prior, including the improper noninformative one. Hence, It implies, asymptotically, the noninformative DP prior does not affect the posterior distribution. Furthermore, Weng (1989) shows such approximation is better than the normal approximation or classical bootstrap approximation in obtaining the posterior distribution of F, because it is also second-order accurate.

Despite its appealing properties above, the BB posterior only includes discrete distributions (with support equal to the sample values) even if the true $F(\cdot)$ is continuous. To correct this continuity problem, histospline smoothing is introduced in Banks (1988). The idea, similar to linear interpolation in the smoothed bootstrap, is to spread probability p evenly between the two closest values in the sample. That is to say, it assigns the posterior Dirichlet mass uniformly across statistically equivalent blocks, formed by distinct values in the data. As n distinct values split a real line into $n + 1$ intervals, we need $n + 1$ probabilities for those intervals. For example, let $X_{n:0}$ and $X_{n:n+1}$ be the lower and upper bound of RV X. We have the posterior $(p_1, \ldots, p_{n+1}) \mid \mathbf{X} = \mathbf{x} \sim \mathrm{Dir}(\mathbf{1})$, where p_k is the probability for interval $(X_{n:k-1},\, X_{n:k}]$

and $\mathbf{1}$ is an $(n+1) \times 1$ vector of ones.

For the continuity correction, p_k is uniformly spread over the interval $(X_{n:k-1},\ X_{n:k}]$ in the sense that the density in the interval is given by $p_k/(X_{n:k} - X_{n:k-1})$. By interpolating within all intervals, the CDF in the Banks (1988) BB is not a step function but an increasing linear spline function with knots at the sample values. In terms of comparing SD, it helps avoid the situation that the SD is decisively rejected by some single extreme value, such as the right endpoint at which an empirical CDF is always 1.

The algorithm for our basic BB inference on $F(\cdot)$ can be summarized as follows.

Step 1. For an ordered sample $\{X_{n:i}\}_{i=1}^{n}$, simulate $\{p_i\}_{i=1}^{n+1}$ from the posterior distribution $\mathrm{Dir}(1, \ldots, 1)$: draw $n+1$ independent RV $\{C_i\}_{i=1}^{n+1}$ from the gamma distribution $\Gamma(1, 1)$, and let $p_k = C_k / \sum_{i=1}^{n+1} C_i$.

Step 2. Construct the random distribution function $T_n(\cdot)$ with the continuity correction suggested by Banks (1988): letting $X_{n:0}$ and $X_{n:n+1}$ be the lower and upper bounds of the support of F,

$$T_n(z) = \begin{cases} \sum\limits_{i=1}^{k} p_i & \text{if} \quad z = X_{n:k} \\ \sum\limits_{i=1}^{k} p_i + \frac{p_{k+1}(z - X_{n:k})}{X_{n:k+1} - X_{n:k}} & \text{if} \quad X_{n:k} < z < X_{n:k+1}. \end{cases}$$

Step 3. Repeat the above steps independently B times to obtain $T_n^1(\cdot), \ldots, T_n^B(\cdot)$.

Step 4. The empirical distribution function of $T_n^1(\cdot), \ldots, T_n^B(\cdot)$ approximates the posterior distribution of $F(\cdot)$ for large B.

1.2.2 Bayesian/frequentist connection

We have shown that the BB method can yield the posterior distribution of CDF $F(\cdot)$ given data $\mathbf{X} = \{X_i\}_{i=1}^{n}$. Now, we want to show its similarity to the frequentist sampling distribution and, more visually, uniform confidence band. Despite the equivalence (as we will show) of the posterior and sampling distributions, and the equivalence of uniform confidence and credible bands, the subtle yet philosophical differences between Bayesian and frequentist perspectives can be extremely important when dealing with SD as they answer essentially different questions.

Suppose $(X_{n:1}, \ldots, X_{n:n})$ are the order statistics of an iid sample $\mathbf{X}$ from a population having a continuous $F(\cdot)$. Wilks (1962, **8.7.1–2**) presents that the sampling distribution of random variables, $F(X_{n:1}), F(X_{n:2}), \ldots, F(X_{n:n})$, is the *ordered* n-variate Dirichlet distribution $\text{Dir}^*(1, \ldots, 1; 1)$, and the marginal distribution of $F(x_{n:k})$ is the beta distribution $\text{Beta}(k, n+1-k)$. It is easier to see the frequentist/Bayesian connection by using Wilks' *coverages* $U_1 = F(X_{n:1}), U_2 = F(X_{n:2}) - F(X_{n:1}), \ldots, U_n = F(X_{n:n}) - F(X_{n:n-1}), U_{n+1} = 1 - F(X_{n:n})$. Such coverages can be understood as the probabilities assigned to the intervals formed by consecutive order statistics. Wilks (1962, **8.7.4**) states that they follow the Dirichlet distribution $\text{Dir}(1, \ldots, 1)$.

Therefore, the frequentist sampling distribution of the coverages in Wilks (1962) is obtained in the same way as their posterior distribution from Banks (1988), specifically from a $(n+1)$-variate $\text{Dir}(1, \ldots, 1)$. The posterior probabilities $\mathbf{p} = (p_1, \ldots, p_{n+1})$ correspond to the coverages $\mathbf{U} = (U_1, \ldots, U_{n+1})$. The differences lie in: the frequentist approach considers the sampling distribution of $\mathbf{U}$ over the possible sample $\{X_{n:k}\}_{k=1}^n$ values conditional on the true unknown value of $\mathbf{p}$, whereas the Bayesian makes probabilistic inference on the unknown true value of $\mathbf{p}$ conditional on the observed values of $\{X_{n:k}\}_{k=1}^n$. Such differences affect their interpretations but do not affect their finite-sample and limiting distributions. Thus, the Banks (1988) posterior distribution and the Wilks (1962) sampling distribution are identical.

The $1 - \alpha$ uniform confidence band for an unknown function $F(\cdot)$ represents the uncertainty in its estimate, say $\hat{F}(\cdot)$, such that it attains simultaneous coverage probability of $1 - \alpha$. When the sampling distribution for $F(X_{n:1}), \ldots, F(X_{n:n})$ is available as above, Aldor-Noiman et al. (2013) provide a computational algorithm to derive a uniform confidence band for $F(\cdot)$. First, for a given pointwise coverage probability $1-\gamma$, the two-sided and equal-tailed confidence interval for $F(X_{n:k})$, $\forall k \in \{1,\ldots,n\}$, is constructed from its marginal distribution Beta$(k, n+1-k)$, with the endpoints being its $\gamma/2$ and $1-\gamma/2$ quantiles. Second, adjust the pointwise coverage level γ such that the confidence intervals for all $\{F(X_{n:k})\}_{k=1}^{n}$ cover the true values simultaneously with $1-\alpha$ coverage probability. This can be done by finding the smallest two-sided p-value for each simulated sample and setting γ to the 100α-percentile over these p-values. Third, form a uniform confidence band for $F(\cdot)$ by interpolating the n confidence intervals. In particular, extend the lower endpoint of the confidence interval at $X_{n:k}$ horizontally toward $X_{n:k+1}$ (and then jump up to its lower endpoint), and extend the upper endpoint of the confidence interval at $X_{n:k+1}$ horizontally toward $X_{n:k}$ (and then jump down to its upper endpoint). The monotonicity of $F(\cdot)$ guarantees the band maintains exact coverage probability.

The uniform confidence band constructed above is also a valid uniform credible band for the BB method in the Banks (1988). A $1-\alpha$ credible band is similarly formed by the set of credible intervals at the order statistics, which have joint $1-\alpha$ credibility. It should be noted that the credible band does not specify the pointwise posterior probability as constant. On the other side, it has been shown that the posterior distribution matches the frequentist sampling distribution in Wilks (1962). Such a sampling distribution is utilized in the first two steps of uniform confidence band construction. The third step does not change the confidence level of the band. Thus, the particular frequentist method provides a uniform credible band for the unknown $F(\cdot)$ in our nonparametric Bayesian problem.

Though the two equivalences are established, their philosophical difference still matters in terms of interpretation. The Bayesian credible band, given the current data, measures (our belief of) the probability that the true F falls within the band.

In contrast, the frequentist confidence band is, when repeated sampling data from population, a measure of how often the constructed band can cover the true F. In short, they answer different questions, though their procedures are superficially identical.

1.2.3 Sampling weights

We consider next the situation where observations are still sampled independently from a population but with different selection probabilities. That is, the sample weights are defined as inverse of selection probabilities. It is easier to extend our basic Bayesian method to account for sample weights than frequentist solutions.[3]

The sample weights, when available, are indispensable in estimating population descriptive statistics such as the CDF. As the National Longitudinal Survey of Youth 1997 (NLSY97)'s technical report puts it, weighting makes the sample representative of the target population. For example, Solon et al. (2015) show the importance of sample weights by using raw data alone in the Panel Study of Income Dynamics (PSID) to estimate the 1967 poverty rate, which is 26%, twice as high as the official measurement by the US census. Therefore, it is almost impossible to obtain unbiased and consistent estimates on descriptive statistics without incorporating sample weights.

One way to use sample weights for inference on F is through the concept of a weighted distribution defined in Rao (1965). One example is to sample the fiber length X: it is more likely to select a longer fiber. That is, the likelihood of a fiber's inclusion in the sample is decided not only by the distribution of X, but the length itself. It makes the recorded length X_i not an observation on X, but on another RV X^w. Therefore the distribution of X^w is called the weighted distribution of the

[3] For example, Kolenikov (2010) discusses variance estimations in complex survey data by three resampling methods, including the frequentist bootstrap. All of them involve repeated calculation of implied weights for each replicate, and method-wise requirements on the number of replications.

original RV X. Formally, for the CDF F of X and its weighted counterpart G, the weighted distribution model can be defined as

$$X_1, \cdots, X_n \mid F \quad \text{are an iid sample from a univariate distribution } G(\cdot \mid F) \tag{1.3}$$

where

$$G(ds \mid F) = w(s)F(ds) \Big/ \int w(s)F(ds) \tag{1.4}$$

$w(s)$ is a known weighting function with $0 < w(s) < \infty$ and has same support as F.

Lo (1993) obtains the posterior distribution of F when the sample can be modeled by the model in (1.3) and (1.4). His idea is to pick a prior that contains knowledge of $w(s)$. Specifically, $w(s)$ are sample weights in surveys and are known before inference. A natural choice of prior is to extend the previous DP prior: the new prior is defined by normalized *weighted* gamma process, in the same manner of DP$(\alpha(\cdot))$ defined in terms of normalized gamma process (Ferguson, 1973, p. 271). In particular, the weighted gamma process is defined by Dykstra and Laud (1981) as $\gamma\big(t; \alpha(\cdot), \beta(\cdot)\big) = \int_0^t \beta(s)Z(ds)$, where $\beta(\cdot)$ is a rate parameter and $Z(ds)$ is a gamma process with independent increments corresponding to shape parameter $\alpha(ds)$. A normalized weighted gamma process, denoted as $\widetilde{\gamma}\big(t; \alpha(\cdot), \beta(\cdot)\big)$, is $\gamma(t)/\gamma(\infty)$. Note DP is a special case of normalized weighted gamma process, i.e., DP$(\alpha(\cdot)) = \widetilde{\gamma}\big(\alpha(\cdot), \mathbf{1}\big)$. Further, Lo (1993) proves, under this improper prior, the posterior distribution of the original F could be first-order approximated by the BB's random distribution of T_n in (1.1).

In practice, only Step 1 of the algorithm in section 1.2.1 needs to be modified for sample weights. Here we illustrate only the major change when considering sample weights. Let $\{W_i\}_{i=1}^n$ be sample weights associated to a sample $\{X_i\}_{i=1}^n$. Instead simulating each Z_i identically from $\Gamma(1, 1)$, we do the following.

Step 1. Simulate C_i from the gamma distribution $\Gamma(1, 1/W_i)$ with rate parameter being the reciprocal of its weight. Note C_{n+1} is used for continuity correction and can be drawn from $\Gamma(1, 1/\bar{W})$ where $\bar{W}$ is the average weight;

1.2.4 Clustering

Clustering arises when arbitrary correlations exist within clusters and independence only holds across clusters. In the frequentist framework, failure to account for the clustering structure can lead to a downward-biased standard error and then over-rejection of the null hypothesis. Cameron et al. (2008) deal with it by resampling, instead of observations at the individual level, the whole set of observations at the cluster level.

Clustering structure is also important in the Bayesian framework. The correlation within cluster poses a huge danger to the validity of BB since it also hinges on the iid assumption. In fact, the number of independent information is indeed the number of clusters for the clustered sample. For example, when investigating the SD relationship between income distributions in two year, a reasonable cluster would be state since people from the same state may be affected by the same economic environment and state-level policies. Thus, the actual number of independent "observations" on income is 50, the number of states. Regarding CDF estimation, we treat each cluster as one "observation" and apply BB to obtain the posterior probability for each cluster. Within each cluster, the probabilities can further be shared by units according to their sample weights.

Consider, specifically, a sample $\{X_{ig}\}$ with G clusters (subscripted by g) and each having N_g observations (subscripted by i); $\{W_{ig}\}$ are associated sample weights. Now Step 1 of the algorithm in section 1.2.1 is modified to simulate posterior probabilities $\{p_{ig}\}$ from the following two steps.

Step 1a. Simulate the cluster-level posterior probability $\{p_g\}_{g=1}^{G+1}$ by $p_g = C_g / \sum_{j=1}^{G+1} C_j$, where C_g is drawn from the gamma distribution $\Gamma(1, 1/\sum_i W_{ig})$ with rate parameter being the reciprocal of the sum of weights within cluster g. Note C_{G+1} is used for continuity correction and can be drawn from $\Gamma(1, 1/\bar{W})$.

Step 1b. The individual-level posterior probability p_{ig} is equal to $p_g W_{ig}/\sum_{j=1}^{N_g} W_{jg}$.

1.3 Nonparametric Bayesian inference on SDj

Here we present nonparametric Bayesian inferential procedures in the context of SD. Section 1.3.1 first characterizes the relevant SD relationships. Then section 1.3.2 shows a Bayesian inferential procedure, obtaining simultaneously the posterior probabilities of all dominance relationships (SD and *non*SD). It also attempts to quantify the differences between two CDFs by a uniform credible band.

1.3.1 Characterization of SDj

Stochastic dominance (SD) provides an unambiguous (partial) stochastic ordering between two RVs or distributions. Its connection to economics has been rigorously studied for poverty and inequality (Ravallion, 1994), and more broadly, for social welfare problems (Deaton, 1997). For example, the income distribution X stochastically dominates Y at the first order, denoted as $X\mathrm{SD}1Y$, iff X has less poverty than Y for any given income level, and X is also preferable for any non-decreasing social utility function. In addition, $X\mathrm{SD}2Y$ is the sufficient and necessary condition for X having higher social welfare for any non-decreasing, concave social utility function. Second-order SD is also equivalent to generalized Lorenz dominance, which implies (among other properties) that the average income $\bar{X}$ is no less than that $\bar{Y}$. Moreover, third-order SD implies higher social welfare when social utility is further restricted to decreasing absolute risk aversion (DARA).

Generally, SD between any two RVs X and Y is defined by their CDFs F_X and F_Y:

1. X SD$_1$ $Y \iff F_X(z) - F_Y(z) \leq 0$ for any $z \in \mathbb{R}$;
2. X SD$_2$ $Y \iff \int_{-\infty}^{z} [F_X(v) - F_Y(v)]\, dv \leq 0$ for any $z \in \mathbb{R}$;
3. X SD$_3$ $Y \iff \int_{-\infty}^{z} \int_{-\infty}^{w} [F_Y(v) - F_X(v)]\, dv\, dw \leq 0$ for any $z \in \mathbb{R}$;

and so on. Davidson and Duclos (2000) introduce a function $D(\cdot)$ to make the characterization of SD neat and convenient. Let $D^1(z) \equiv F(z)$ and

$$D^j(z) \equiv \int_{-\infty}^{z} D^{(j-1)}(v)\, dv = \frac{1}{(j-1)!} \int_{\infty}^{z} (z - v)^{(j-1)}\, dF(v), \quad j = 1, 2, \ldots. \tag{1.5}$$

Therefore, distribution X is said to dominate distribution Y stochastically at order j if and only if $D_X^j(z) - D_Y^j(z) \leq 0$ for any $z \in \mathbb{R}$.

Although the SD relationships can be hypothesized directly as the conditions above, specifically over the whole support like $z \in \mathbb{R}$, there are few papers doing so. The main reason, as Davidson and Duclos (2013) point out, is that the whole support range is not statistically feasible because of too little reliable information on income and other variables in the tails. For example, it still rejects the null of SD when one CDF curve is everywhere lower than another except at some endpoints. Such rejections are often awkward, as the extreme values may be due to measurement error or sampling error. Therefore, we consider the SD relationships over a restricted support of incomes $z \in [\underset{\sim}{z}, \bar{z}]$, where $\underset{\sim}{z} = \max\{\min(x), \min(y)\}$ and $\bar{z} = \min\{\max(x), \max(y)\}$. The general hypotheses for testing stochastic dominance of order j are formed as

$$\begin{array}{lll} H_0^j: & D_Y^j(z) - D_X^j(z) \leq 0 & \text{for all } z \in [\,\underset{\sim}{z}, \bar{z}\,] \\ H_1^j: & D_X^j(z) - D_Y^j(z) \leq 0 & \text{for all } z \in [\,\underset{\sim}{z}, \bar{z}\,] \\ H_2^j: & \begin{cases} D_X^j(z_1) - D_Y^j(z_1) < 0 \\ D_Y^j(z_2) - D_X^j(z_2) < 0 \end{cases} & \text{for some } z_1, z_2 \in [\,\underset{\sim}{z}, \bar{z}\,] \end{array} \tag{1.6}$$

The three hypotheses H_0^j, H_1^j, and H_2^j respectively correspond to $Y\text{SD}jX$, $X\text{SD}jY$, and neither X nor Y dominating at order j.

1.3.2 Nonparametric Bayesian inferential procedure

The Bayesian approach can summarize the data about the relationship between two distributions in two useful ways. First, the hypothesized statements in (1.6) can be assigned posterior probabilities. Second, a uniform credible band for the difference between the two CDFs can be computed. Additionally, posterior probabilities can be converted to a binary decision such as reject/accept by minimizing posterior expected loss based on a chosen loss function.

If X and Y are independent, then the posterior probabilities of all possible dominance relationships at order j can be simulated based on $D^j(\cdot)$ in (1.5) and the posterior distributions of $F_X(\cdot)$ and $F_Y(\cdot)$ in section 1.2. For example, consider second-order SD given samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, where n and m may differ. Let $\mathbf{Z} = \{Z_{l:i}\}_{i=1}^l$ be the order statistics in the union of distinct values of the two samples, so $l \leq n+m$.

The posterior probabilities for SD2 are simulated as follows.

Step 1. Simulate $\mathbf{p}^\mathbf{X}$ from the posterior distribution $\text{Dir}(1, \ldots, 1)$ derived in section 1.2.1 and construct $D_X^2(z)$ for all $z \in \mathbf{Z}$. Similarly, draw $\mathbf{p}^\mathbf{Y}$ and construct $D_Y^2(z)$ for sample $\mathbf{Y}$.

Step 2. Let $\Delta_2(z) \equiv D_X^2(z) - D_Y^2(z)$.

Step 3. Repeat Step 1 and Step 2 R times. For large enough R, the posterior probabilities of dominance relationships are well-approximated by

$$\begin{aligned} \mathrm{P}(H_0^2\colon X\mathrm{SD}2Y \mid \mathbf{X},\mathbf{Y}) &= \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{\Delta_2(z) \le 0, \forall z \in \mathbf{Z}\} \\ \mathrm{P}(H_1^2\colon Y\mathrm{SD}2X \mid \mathbf{X},\mathbf{Y}) &= \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{\Delta_2(z) \ge 0, \forall z \in \mathbf{Z}\} \\ \mathrm{P}(non\mathrm{SD}_2 \mid \mathbf{x},\mathbf{y}) &= 1 - \mathrm{P}(H_0^2 \mid \mathbf{X},\mathbf{Y}) - \mathrm{P}(H_1^2 \mid \mathbf{X},\mathbf{Y}). \end{aligned}$$

The Bayesian approach is quite appealing because the notion of the probability of a hypothesis, such as $\mathrm{P}(H_0 \mid \mathbf{X},\mathbf{Y})$, can only be defined in the Bayesian way. In particular, $X\mathrm{SD}jY$ is an unknown population relationship and to be tested from data. Its posterior probability, given the data, provides a probabilistic measurement on the degree of belief on this fixed and hypothesized statement. What's more, three possible situations in (1.6) are systematically inferred by treating them symmetric. It is a striking difference from the frequentist approach, which places asymmetric roles in null and alternative hypotheses.

Moreover, a uniform credible band[4] of $\Delta_j(\cdot)$ can be derived from the simulations. Specifically, replace Step 3 with the following.

Step 3. Repeat Step 1 and Step 2 many times to approximate the joint posterior distribution of $\Delta_j(z)$ for all $z \in \mathbf{Z}$.

Step 4. Find the constant width w such that the band $[\overline{\Delta_j(\cdot)} - w, \overline{\Delta_j(\cdot)} + w]$ contains $\Delta_j(\cdot)$ with posterior probability $1-\alpha$, where $\overline{\Delta_j(\cdot)}$ is the posterior mean.

Finally, it is often pragmatic to require a deterministic conclusion on the hypotheses by restricting the decision space to {Accept, Reject}, or equivalently $\{0,1\}$. To this end, a loss function is needed. For instance, the 0–1 loss, formalized by Neyman and Pearson, incurs the penalty (loss) of one if the decision is wrong and zero otherwise. It is a qualitative loss in the sense that it does not differentiate the two types of

[4]The notion is discussed in section 1.2.2.

error and thus fails to weigh the importance of the null hypothesis in the loss. More quantitatively, a weighted 0–1 loss takes value $1-\alpha$ for type I error, α for type II error, and zero otherwise. To minimize its posterior expected loss (PEL), the optimal decision is given in proposition 1.

Proposition 1. *Under the weighted 0–1 loss above, reject H_0 if $\mathrm{P}(H_0 \mid \mathbf{X}) \leq \alpha$; otherwise, accept H_0.*

Intuitively, the PEL is minimized when the posterior expected loss from false rejections is equal to the posterior expected loss from false acceptance, that is, $(1-\alpha)\,\mathrm{P}(H_0 \mid \mathbf{X}) = \alpha(1-\mathrm{P}(H_0 \mid \mathbf{X}))$. The ratio of $(1-\alpha)/\alpha$ in the loss reveals the relative importance of the type I error against the type II error. The larger it is, the smaller the posterior probability of H_0 needs to be for H_0 to be rejected.

1.4 Simulations

It is instructive to place frequentist and Bayesian inference together under scrutiny. In this section, we compare our Bayesian bootstrap inference (BB) with the Kolmogorov–Smirnov (KS) type bootstrap tests in Barrett and Donald (2003) (hereinafter BD03) and the empirical likelihood ratio (ELR) tests in Davidson and Duclos (2013) (hereinafter DD13). Such comparisons are measured using two relevant criteria: the first one is the frequentist risk (i.e., the expected loss averaged over random data only); the second is the Bayes risk (i.e., the expected loss averaged over the data and unknown parameters). The latter can also be understood as the frequentist risk averaged over the parameter space against to its prior distribution.[5]

The major conclusions are summarized here. Comparing to frequentist inference, BB inference on SD is better when measured by Bayes risk, while it is worse when measured by frequentist risk. In the frequentist sense, it is anti-conservative for

[5]Details on the two criteria connections can be found in Robert (2007).

the null of SD, but conservative for the null of *non*SD. These results agree with the findings in Kaplan and Zhuo (2017). Here, they imply that the Bayesian inference may not necessarily have the frequentist properties in joint hypothesis testing contexts.[6] However, when we account for uncertainties in parameters and use Bayes risk, the BB outperforms its frequentist counterparts.

1.4.1 Frequentist risk of tests of the SD null

Our first simulation aims to inspect the null of SD by replicating the Monte Carlo experiment in BD03. Though all five cases in their paper are investigated, only the first and fourth cases are illustrated here in order to focus on size control and power. In particular, the two DGPs assume iid sampling with X and Y independent, with distributions specified as follows.

Case 1: $X_i \sim \ln \mathrm{N}(0.85, 0.6^2)$ and $Y_i \sim \ln \mathrm{N}(0.85, 0.6^2)$, where $\ln \mathrm{N}(\mu, \sigma^2)$ is a log-normal distribution with location μ and scale σ.

Case 2: X_i is the same as in Case 1, and Y has the following mixture distribution: $Y_i \sim \ln \mathrm{N}(0.8, 0.5^2)$ w.p. 0.9 and $Y_i \sim \ln \mathrm{N}(0.9, 0.9^2)$ w.p. 0.1.

Therefore, all SD hypotheses are true for Case 1 and false for Case 2.

We consider stochastic dominance of X by Y up to order 3. The Bayesian decision is, as shown in proposition 1, to reject H_0^j if $\mathrm{P}(H_0^j \mid \mathbf{X}, \mathbf{Y}) \leq \alpha$. The posterior probability $\mathrm{P}(H_0^j \mid \mathbf{X}, \mathbf{Y})$ is calculated by the algorithm in section 1.3.2. The decision rule for the KS-type test is to reject H_0^j if $\hat{p}_j \leq \alpha$, where $\hat{p}_j$ is the p-value for the corresponding test statistic. The p-value is obtained by the second bootstrap method detailed in BD03.

[6] In fact, the BB inference shows asymptotically correct size control when testing a general one-sided hypothesis of a location parameter. For example, imagine one population CDF touches another at a single point, and otherwise stays on its left side. This equality at the lone contact point is crucial in inferring the population SD1.

A total of 1000 Monte Carlo replications are performed for each case. We consider two sample sizes of $N_x = N_y = 50$ and $N_x = N_y = 500$. Table 1.1 reports the rejection rates for BB and KS tests at a significance level of 5%.

Table 1.1: Rejection probability, $\alpha = 5\%$.

		Case 1		Case 2	
H_0	$N_x(= N_y)$	Bayesian	frequentist	Bayesian	frequentist
YSD1X	50	75.1 %	3.3 %	84.2 %	7.1 %
YSD2X	50	31.3 %	6.0 %	33.3 %	13.8 %
YSD3X	50	25.0 %	5.8 %	24.5 %	13.7 %
YSD1X	500	87.8 %	5.0 %	99.3 %	46.9 %
YSD2X	500	29.3 %	4.7 %	63.8 %	45.7 %
YSD3X	500	23.3 %	5.2 %	42.8 %	43.9 %

Note: all H_0 in Case 1 are true, while all H_0 in Case 2 are false.

Case 1 focuses on size control. The BB test fails to control its rejection probability (RP) at the nominal level of 5% for all orders (e.g., it rejects the SD1 null in 75.1% of repetitions when $N_x = 50$). However, Kline (2011) argues that this behavior is more appropriate and intuitive than controlling size, in an example equivalent to testing discrete SD. Kaplan and Zhuo (2017) try to explain the phenomenon geometrically, due to the DGP in Case 1 being not only on the boundary of the null hypothesis space, but at a very (very) sharp corner of it. For example, if the null of SD1 contains k inequalities, say $v_1 = F_x(z_1) - F_y(z_1), \ldots, v_k = F_x(z_k) - F_y(z_k)$, then the orthant, or hyperoctant, of Euclidean space $\mathbb{R}^k = (v_1, \ldots, v_k)$ satisfying H_0 is convex. The volume of the corresponding hypercube makes up only $1/2^k$ proportion in $\mathbb{R}^k$. When k is large, its volume is so small that the likelihood of SD1 cannot be substantial without strong evidence favoring it. Such explanations are also consistent with 1) higher-order SD suggests a larger volume of the hypercube satisfying SD and thus larger posterior probability and lower RP; 2) larger sample size may mean more inequalities and smaller proportion of the hypercube, and thus lower posterior probability and higher RP. It should be mentioned that the high RP for the Bayesian inference is also observed by Kline (2011), which claims the frequentist test is too conservative when testing one-sided multivariate hypotheses.

Case 2 checks the power of the tests. Overall, BB inference makes type II error less often and has better power. It dominates the KS test for all orders when sample size is small ($N_x = 50$). Though the KS test sees a rapid increase in power for a larger sample, it still makes more type II errors than BB. Along with Case 1, this shows that BB inference is logically coherent: the higher-order SD null always has a lower RP since it is less restrictive. In contrast, the KS test does not have this property.

1.4.2 Bayes risk of the SD null

We now turn to a one-sample problem and investigate SD relationships with a known CDF. This Monte Carlo study shows (Bayes risk) measures by which BB inference is the better choice. The DGPs are as follows. Now, X is known to be standard uniformly distributed, i.e., $F_X(z) = z$ for $z \in [0, 1]$. The RV Y is also defined separately over h equal sub-intervals in $[0, 1]$, with $F_Y(\cdot)$ linearly increasing on each segment. That is, the distribution function of Y is fully specified by $(F_Y^{(1)}, \ldots, F_Y^{(h-1)})$, i.e., F_Y at the $k - 1$ kink points.

To calculate Bayes risk, we treat $\mathbf{F} = (F_Y^{(1)}, \ldots, F_Y^{(h-1)})$ as random. A total of 5000 Monte Carlo replications are performed. Each replication first draws $\mathbf{F}$ from the ordered $(h-1)$-variate Dirichlet distribution (see definition in Wilks, 1962, p. 236). Second, given this realized $\mathbf{F}$ and corresponding $F_Y(\cdot)$, a data sample of size N_y is drawn from this realized distribution function.[7]

Table 1.2 compares Bayes risk of the BB and KS-type hypothesis tests. We use the weighted 0–1 loss function with $\alpha = 0.05$. BB has smaller Bayes risk than KS test in almost every situation. Both of them have higher Bayes risk for second-order SD.

[7]For practical purpose, we also assume the lower and upper boundary uniformly over $[-0.01, 0.01]$ and $[0.99, 1.01]$.

Table 1.2: Bayes risk (measured at one thousandth)

		$h=4$		$h=8$	
H_0	N_y	Bayesian	frequentist	Bayesian	frequentist
YSD1Unif$(0,1)$	50	10.59	25.48	8.86	27.70
YSD2Unif$(0,1)$	50	28.66	28.19	26.36	30.22
YSD1Unif$(0,1)$	500	2.32	16.18	2.02	11.26
YSD2Unif$(0,1)$	500	13.10	21.02	11.37	19.84

1. Loss function takes value 0.95 for type I error, 0.05 for type II error, 0 otherwise.
2. $h-1$ is number of random parameters in $F_Y(\cdot)$.

In the sample size of 50, the Bayes risk for BB decreases when h increases to 8 from 4. In contrast, the KS test suffers from more parameter uncertainties and increases its Bayes risk. It suggests BB inference may be better when we have small sample and more uncertainty. Naturally, the larger sample size helps both to reduce their Bayes risks. But BB inference still remains a better choice.

1.4.3 The null of *non*SD

The last Monte Carlo study replicates DD13, whose focus is on the null of *non*SD. For a frequentist test, if one wants to seek a conclusion of dominance, it is better to posit *non*SD as the null hypothesis: it is more conclusive to reject non-dominance than to fail to reject dominance. But such a null hypothesis is very statistically demanding. This simulation study not only shows the coherence and flexibility of BB inference on SD and *non*SD, but gives another chance to look at its frequentist properties.

Two samples of X and Y are independent. X is drawn from the standard uniform distribution and Y is defined over eight equal sub-intervals in $[0,1]$, with $F_Y(\cdot)$ continuous and linearly increasing on each segment. The values of $F_Y(\cdot)$ evaluated at the upper limit of each segment are 0.03, 0.13, 0.20, 0.50, 0.57, 0.70, and 1.00. To sum up, $F_Y(\cdot)$ stays below $F_X(\cdot)$ everywhere, except at $z=0.5$, where $F_X(0.5)=F_Y(0.5)=0.5$. This DGP is on the boundary between SD and *non*SD, so rejection probabilities may be interpreted as type I error rates for either null.

In table 1.3, we give the rejection rates of various tests under different sample sizes and nominal level α for a total of 1000 Monte Carlo replications.[8] We compare our BB inference with DD13's empirical likelihood ratio (ELR) test for the null of *non*SD, and with BD03's KS-type test for the null of SD.

Table 1.3: Rejection probability, nominal level α.

		$\alpha = 1\%$		$\alpha = 5\%$	
H_0	N_x	Bayesian	frequentist	Bayesian	frequentist
Y *non*SD_1 X	32	0.1 %	0.1 %	0.6 %	1.8 %
Y *non*SD_1 X	128	0.1 %	0.6 %	1.5 %	4.9 %
Y *non*SD_1 X	512	0.3 %	1.0 %	2.0 %	4.9 %
Y *non*SD_1 X	1024	0.3 %	1.0 %	3.0 %	5.1 %
YSD1X	32	4.0 %	0.0 %	15.3 %	1.4 %
YSD1X	128	3.1 %	0.0 %	11.6 %	1.1 %
YSD1X	512	2.7 %	0.0 %	7.5 %	1.0 %
YSD1X	1024	1.3 %	0.0 %	6.8 %	0.8 %

Note: DGPs and methods as described in the text. "Frequentist" uses the ELR test in DD13 for the *non*SD_1 null, and the KS test in BD03 for SD1.

As is evident in table 1.3, BB inference under-rejects the *non*SD_1 null, but over-rejects the SD1 null. For a given sample size, these facts could also be explained by the "convexity" of the null hypothesis. If H_0, for example, contains k inequalities, the volume of the SD1 hypercube makes up only $1/2^k$ proportion, while the volume of the *non*SD_1 hypercube takes the rest proportion. Therefore, it is more likely to satisfy the null of *non*SD_1. That is, RP is higher for SD1 than for *non*SD_1. Besides, as the sample size grows, RP is converging to the nominal level α, and the difference between the two nulls is closing. It is consistent with the findings of Casella and Berger (1987) in the one inequality case: the Bayesian posterior probability can equal the frequentist p-value. In this DGP, only one inequality at $z = 0.5$ would be binding at zero and the rest would not in large samples.

[8] According to DD13, the samples drawn from X are of sizes $N_x = 32, 128, 512, 1024$. Correspondingly, $N_y = 19, 91, 379, and 763$, the rule being $N_y = 0.75N_x - 5$.

On the frequentist side, they are conservative for both SD and *non*SD nulls in small samples. Asymptotically, only the ELR test for *non*SD has correct size, while the KS test becomes more conservative. Therefore, the differences between two opposite nulls are widening, rather than closing. The ELR test is asymptotically size-correct for the null of *non*SD, though it under-rejects the null of *non*SD in the small sample. This observation could possibly be explained by that it is easier to fail to reject SD than to reject SD. Several questions could be raised, such as does the KS test fail to control the size for this specific question? Is the KS test a still consistent test for the case where we have, instead $F_Y(0.5) = F_X(0.5)$ like here, $F_Y(0.5) = F_x(0.5) + \epsilon$ and $\epsilon \to 0^+$?

1.5 Empirical applications

In this section we consider the Bayesian bootstrap (BB) test in the context of an empirical example. The data we use comes from the Panel Study of Income Dynamics (PSID) for years 1997 and 2013. We consider comparisons of the income distributions[9] in 1997 and 2013 without and with sample weights. In table 1.4 we provide some basic descriptive statistics for these data. In addition, in fig. 1.1A we plot the emprical CDFs (ECDF) for the income data with the 1997 distribution being the solid line. Figure 1.1B contains the difference between the 1997 and 2013 ECDFs against income values and gives a much clearer picture. Similarly, Figures 1.2A and 1.2B show the weighted ECDFs and their difference when sample weights are used to guarantee its representativeness of the populations. As indicated by the figures, the difference between these distributions is quite erratic even though the distributions themselves are close. The maximum absolute difference of the two CDFs, around 2%,

[9] The incomes are defined as per capita incomes, that is, the average real income earned per person in a household. The individual weight in a household is 1 for adult and 0.5 for children. All incomes are measured in the dollar of the year 2000.

is relatively large if we recall CDFs range from 0% to 100%. Moreover, Figures 1.1B and 1.2B also give one an idea of the importance of considering sample weights: the weighted difference is dramatically different from the unweighted one. Therefore, the conclusions based on the raw ECDFs may be inaccurate, even misleading.

Table 1.4: Descriptive statistics of real per capita income (in year 2000 dollars)

		Raw			Weighted		
Year	Sample	Mean	Std. Dev.	Median	Mean	Std. Dev.	Median
1997	6747	22 880	24 229	17 182	26 149	26 635	19 956
2013	8907	23 468	29 378	17 186	27 724	33 393	20 697

Note: weighted mean $\tilde{E} = \sum wx / \sum w$; and Std. Dev=$\sqrt{\sum w(x-\tilde{E})^2/(\sum w - 1)}$.

In table 1.5 we present p-values for the bootstrap KS test in BD03[10] and the posterior probabilities for our BB method in this paper, for the 1997/2013 income distribution comparison. The left panel labeled "1997 SDj 2013" presents the results for testing whether the income distribution in year 1997 stochastically dominates the income in 2013 at order j, while the other panel tests the opposite hypothesis.[11]

We have the frequentist and Bayesian results for the raw income in table 1.5. There is not much agreement between the two tests because they are essentially answering two different questions. The Bayesian framework tries to gauge the probability of H_0 conditional on data; the frequentist one measures how well data support the condition/assumption of H_0 being true. In particular, our BB tests suggest there is essentially zero chance that 1997 dominates 2013 at the first order, but the possibilities of SD2 and SD3 rise to 5.0% and 14.8% respectively. For the converse hypotheses, the Bayesian interpretation is that it is essentially impossible that 2013 dominates 1997 for any order up to 3. According to proposition 1, BB tests indicate that, when the loss function takes value 0.95 for Type I error and value 0.05 for Type II error, neither

[10]The p-value is calculated by the code from the authors' homepages.

[11]For example, the null hypothesis for the $j=1$ column of the panel labeled "1997 SDj 2013" is that $D_{1997}^{j=1}(z) - D_{2013}^{j=1}(z) \leq 0$ for any z, i.e., the CDF in 1997 is less than or equal to that in 2013.

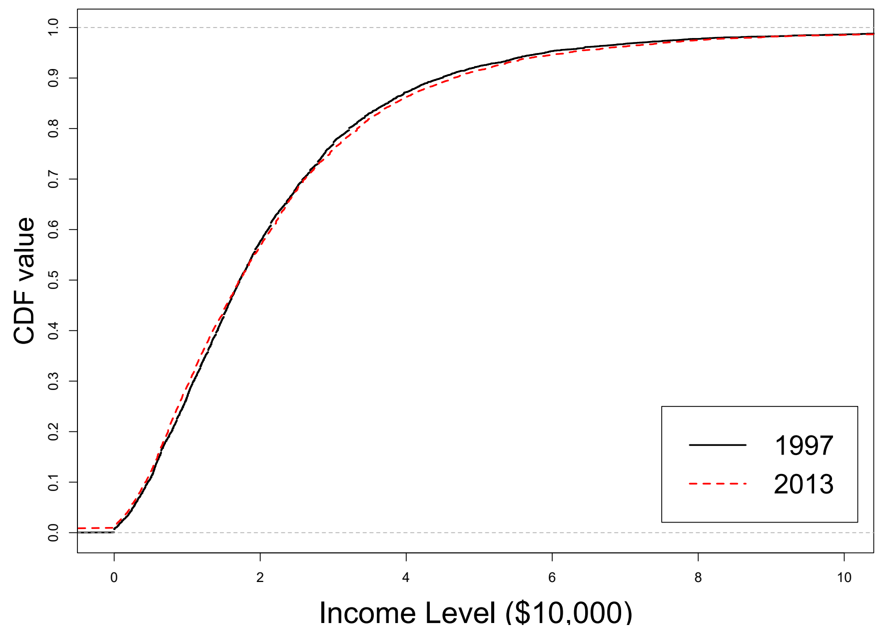


Figure 1.1A: Per capita real income CDF

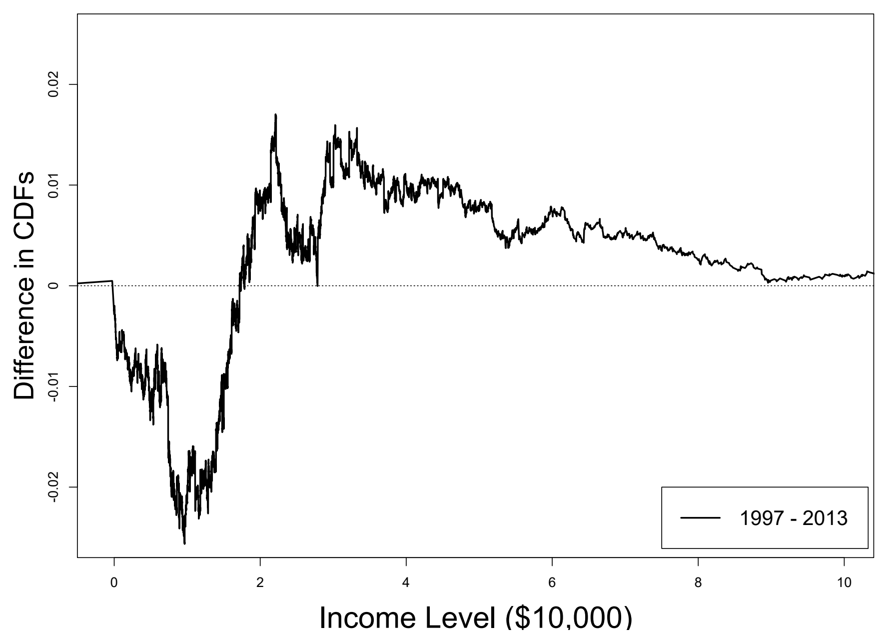


Figure 1.1B: Per capita real income CDF difference

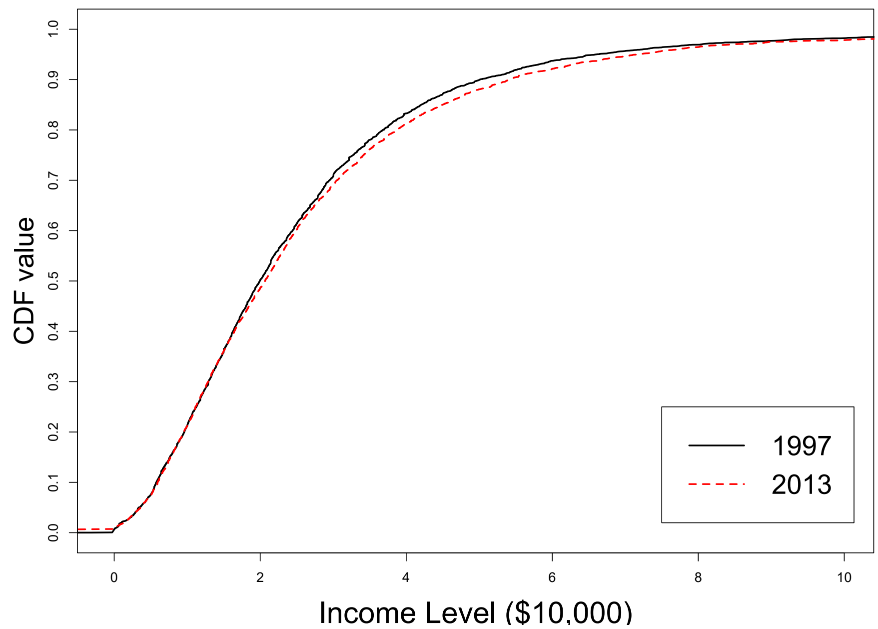


Figure 1.2A: Per capita real income *weights-adjusted* CDF

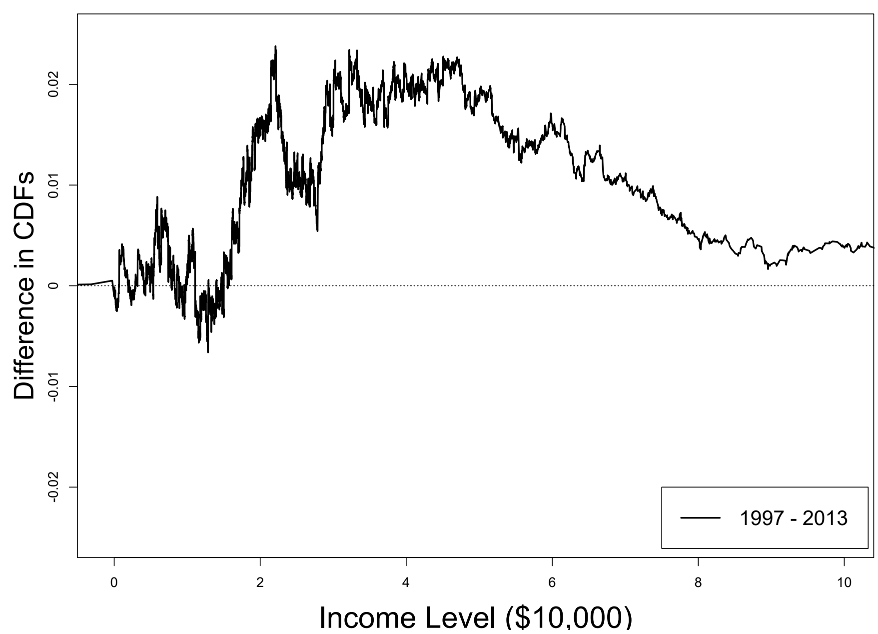


Figure 1.2B: Per capita real income *weights-adjusted* CDF difference

distribution dominates the other in a first-order or second-order sense, and one can accept that the 2013 distribution is dominated by 1997 in a third-order sense. On the other hand, the KS tests suggest one can only reject that the 2013 distribution first-order dominates 1997 at the 5% significance level; other than that, one cannot reject the rest of the stochastic dominance relationships between the two income distributions. Regarding the binary decision (reject/not reject) on the hypothesized stochastic orderings between two income distributions, the BB test rejects more than the KS test does, with the same rejection being made for H_0: 2013SD11997. But it is well-known that non-rejection in frequentist tests is often inconclusive, especially when power is low. A more reasonable approach is to compare the p-value and posterior probability. The KS test favors the evidence when the 2013 distribution is assumed to dominate 1997 for higher orders, with their $p-$values (38.8% and 54.5%) being much higher than those under the converse hypotheses (10.4% and 13.8%). In contrast, the BB test yields higher chances of the latter hypotheses being true after observing the data. To conclude this paragraph, we urge readers to explore fig. 1.1B when trying to understand the divergences between two tests. If, for example, 1997 SD1 2013 is true in the population, the CDFs' difference curve should be negative and stay below the zero all the time.[12] Our BB test examines literally this fact in the difference's posterior distribution, while the KS test focuses on the most positive distance and gives the size protection for the peak of curve.

Sample weights reverse most conclusions above from the Bayesian analysis.[13] In the last row of table 1.5, the BB tests imply that there are zero possibilities that 1997 dominates 2013 at the first two orders, and as slim as 0.2% chance for the third order. For the converse hypotheses, the BB tests indicate there is a 47.4% chance that the 2013 distribution dominates 1997 at either the second or third order. It is

[12]Similarly, the SD2 can also be expressed in terms of the integral of the difference curve being negative.

[13]The bootstrap-version KS test does not consider sample weights partly since it is relatively hard to re-weight during resampling.

Table 1.5: Stochastic dominance in PSID.

Include weights	Method	1997 SDj 2013			2013 SDj 1997		
		$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
No	BB	0.0 %	12.1 %	25.1 %	0.0 %	0.01 %	0.2 %
No	KS	10.8 %	10.4 %	13.8 %	0.6 %	38.8 %	54.5 %
Yes	BB	0.0 %	0.0 %	0.2 %	0.0 %	47.4 %	47.4 %

1. "X SD Y" means the income distribution in year X stochastically dominates Y.
2. "KS" gives the p-value; "BB" gives the posterior probability.

easier to perceive these changes with the help of fig. 1.2B. For example, most of the difference curve stays above zero, except some fluctuations around zero in a short income range below 18 000. It thus is somehow tempting to infer that 2013 SD 1997. But the weighted median income in 1997 is 19 956, which means that the sample size of incomes within the range is close to half of the total. The BB test quantifies these observations by giving 0% posterior probability for 2013 SD1 1997 and 47.4% for 2013 SD2 1997. What's more, it implies there is a 47.7% probability that the 2013 income distribution should be preferred when the social utility function is non-satiated and risk-averse.

1.6 Conclusion and extensions

In this paper, we have studied first-order and higher-order stochastic dominance between two income distributions X and Y in a nonparametric Bayesian model. We have used the smoothed Bayesian bootstrap of Banks (1988) to obtain the posterior probabilities of three hypotheses, for any order of stochastic dominance: (1) X dominates Y; (2) Y dominates X; and (3) neither distribution is dominant. Although the corresponding hypothesis test we consider does not attain frequentist size control, it often has lower Bayes risk (also a frequentist measure) than frequentist tests. This approach has several other advantages, including coherent probabilistic measurement of the three dominance hypotheses, robustness against parametric misspecification of the distribution, minimal impact of prior information, computational efficiency, and incorporation of sample weights and clustered errors.

Future work may proceed in several directions. First, the methods in this paper may be applied to conditional distributions (under unconfoundedness) and regression discontinuity designs, using results from Canay and Kamat (2017) as in Goldman and Kaplan (2017, §6). Second, it would be interesting to determine the prior (or perhaps loss function) to achieve frequentist size control for the Bayesian test that minimizes posterior expected loss of the accept/reject decision. However, usually "probability matching priors" are to ensure correct coverage probability of credible sets, which is already true here; it is instead the particular shape of the stochastic dominance hypotheses that cause Bayesian and frequentist conclusions to diverge. Third, in ongoing work, we are applying our methods to a portfolio choice problem using a probabilistic version of second-order stochastic dominance efficiency.

Chapter 2

Inferring health inequality from ordinal data

2.1 Introduction

Ordinal data are one of the most popular measures of individual health. These data are often collected as responses to a generic question such as, "Would you say your health in general is excellent, very good, good, fair, or poor?" and hence called self-reported health status (SRHS). SRHS is available in many large-scale surveys and has a considerable predictive power for health-related objectives, such as mortality. On the other hand, its ordinal nature limits its role in characterizing the distribution of "true" health.

Our interest here is to interpret *properly* the distributional information carried by SRHS data. In particular, we are concerned with the following two effects on the health distribution: 1) the age effect, i.e., how health inequality (dispersion) evolves with age; 2) the cohort effect, i.e., how health inequality is changing across generations.

When using SRHS for health inequality, the main problem is that many measures of inequality are mean-based and hence not well defined with ordinal data. Generally, one can either assume a cardinal variable underlying SRHS and then study latent inequality, or else redefine the concept of inequality using quantiles. For the first approach, Deaton and Paxson (1998) specify a discrete health variable and find that its variance increases with age. One can also assume a continuous latent variable with known parametric distribution. For example, Wagstaff and Van Doorslaer (1994) assume a log-normal latent health variable and then transform SRHS into a discrete variable through threshold points.[1] This latent approach is appealing since one can employ many statistical inequality tools for "true" health. However, it may be plagued with the choice of latent distribution and, more broadly, misspecification. Different specifications may yield different inequality rankings. Allison and Foster (2004), for instance, find that mean-based inequality rankings may be highly sensitive to re-scaling of ordinal data. Therefore, they propose a considerably different methodology, which defines inequality by using the median as the reference point. Specificallly, an ordinal health distribution F displays more inequality than another distribution G if F is obtained from G via a sequence of *median-preserving spreads*. Though it serves a similiar goal as the cardinal concept of "mean-preserving spread," it actually reworks first-order *stochastic dominance* (SD) and hence relies only on cumulative proportions of ordinal data. To the best of our knowledge, the current health literature does not further provide inferences on these partial orderings. Instead, median-based indices are proposed as alternatives to overcome statistical difficulties in inferring median preserving spread or SD.[2] These indices introduce ad hoc parameters and may lack ethical robustness.[3]

Another problem is SRHS itself as a health measure, which is subject to reporting

[1]See more examples: Cutler and Richardson (1997) and Groot (2000) for ordered probit model, Contoyannis et al. (2004) for dynamic panel ordered probit model, van Doorslaer and Jones (2003) for interval regression, and Jürges (2007) and Bago d'Uva et al. (2008) for generalized models.

[2]Examples of indices: Abul Naga and Yalcin (2008), Jones et al. (2011), and Kobus and Miłoś (2012).

[3]See arguments in Davidson and Duclos (2000).

biases. Some reporting bias need not be a major concern providing it is random.[4] However, Currie and Madrian (1999) suggest there also exists non-random measurement error in SRHS. One example is heterogeneity in reporting behavior. Lindeboom and van Doorslaer (2004) show that individuals with the same "true" health may use systematically different threshold levels when reporting SRHS.[5] They further classify such reporting heterogeneity into two types, based on whether the thresholds move in a parallel way ("index shift") or not ("cut-point shift"). Empirically, they present some evidence of both shifts for age and sex in the Canadian National Population Health Survey, where the index shift is much more significant. Hernández-Quevedo et al. (2005) also find evidence of an index shift in threshold in the British Household Panel Survey, but little evidence of the cut-point shift. On the other hand, the Panel Study of Income Dynamics (PSID) reveals that when respondents are further required to rate health in their youth, the youth SRHS distribution does not vary significantly across age cohorts. This suggests no systematic heterogeneity in reporting behavior. Overall, evidence of reporting bias is mixed.

We illustrate the limitations of SRHS in inferring health inequality through two important empirical questions. First, we study health inequality over the life cycle. This dynamic is key to understanding individual choices regarding working, saving, and retirement, and thus forming public policies concerning health care, financing, and pensions. Arguably, the most predominant hypothesis is that the dispersion of health grows with age. Deaton and Paxson (1998) provide supporting but shaky evidence since SRHS is contentiously assumed a quantitative variable. Van Kippersluis et al. (2009) use external information to scale SRHS and do not consistently find supporting evidence among European countries. Second, we focus on inter-cohort trends

[4]Investigating (random) measurement error in SRHS, Crossley and Kennedy (2002) find that when SRHS is asked twice within a survey, 28% of people change their response, though only 3% of changes are more than one category.

[5]Their "true" health is measured by a continuous variable called the *McMaster Health Utility Index Mark 3 (HUI3)*.

in health inequality for the elderly. Their health not only reflects consequences of health policies or changes in the health system in the past, but predicts the future use of health services. The current literature focuses mainly on either how persistent health inequality is after retirement (Heiss, 2011; Heiss et al., 2014) or generational differences of health inequality for the young (van Kippersluis et al., 2009).

We contribute to the literature in two ways. First, we provide direct inference on median preserving spread or SD, which are formulated as sets of inequality constraints. Such joint hypothesis testing has recently received great attention in the econometrics literature. In particular, we develop a frequentist test based on the refined moment selection (RMS) procedure in Andrews and Barwick (2012), as well as Bayesian inference using the Dirichlet–multinomial model. Second, both latent and quantile approaches are placed together for comparison. We also comment on connections we perceive under the assumption of fixed thresholds.

The remainder of the paper is in three sections. Section 2.2 presents latent health and quantile approaches used for comparing two health distributions. We further discuss how the particular assumption of stable thresholds links both approaches together. Section 2.3 provides various applicable statistical inferences. Their algorithms are also provided. Section 2.4 presents two age-specific examples using SRHS data from the Current Population Survey (CPS). Section 2.5 concludes the paper.

2.2 Concepts

Let $X \sim F$, a discrete distribution with the support being k ordered categories (from poor to excellent), and let $\mathbf{p} \equiv (p_1, \ldots, p_k)$ be the associated population probabilities. Similarly, G and $\mathbf{q} \equiv (q_1, \ldots, q_k)$ are defined for Y. We now describe different concepts helpful for characterizing and learning about inequality.

2.2.1 Quantile approach: first-order stochastic dominance

First-order SD is a scaling-robust ranking for ordinal random variables, which relies only on their cumulative probabilities.

Definition 1. Given any two ordinal random variables, X and Y, we say that X *first-order stochastically dominates* Y, written $X \text{ SD}_1 \, Y$, if

$$F_j \leq G_j, \quad \forall\, j = 1, \ldots, k, \tag{2.1}$$

where $F_j \equiv \sum_{i=1}^{j} p_i$ is the cumulative share of the population in the first (bottom) j categories of X, and G_j is the analogous quantity for Y.

In the context of health distributions, the inequality $F_j \leq G_j$ ensures that, for people rating their health at category j and better, the cohort Y has a higher percentage of its population than X.

Though second-order SD is equivalent to general Lorenz dominance for income inequality, higher-order SD here is of little help because it requires cardinal value assignments for ordinal data and thus becomes sensitive to arbitrary numerical scaling.

2.2.2 Quantile approach: median-preserving spread

Allison and Foster (2004) propose to use the median preserving spread relationship for health inequality as an alternative to (general) Lorenz dominance. They characterize inequality comparisons by using the median as the reference point and modifying first-order SD as follows.

Definition 2. Given any two ordinal random variables, X and Y, we write X MD Y to denote that X dominates Y in the sense of Y being a median preserving spread of X if

1. the median in X and Y remains in the same category m;

2. for all $j < m$, $F_j \le G_j$;

3. for all $j \ge m$, $F_j \ge G_j$;

where, as in definition 1, $F_j \equiv \sum_{i=1}^{j} p_i$ and $G_j \equiv \sum_{i=1}^{j} q_i$.

That is, Y is a median preserving spread of X if X SD$_1$ Y below the median while Y SD$_1$ X for the median and above. It suggests that X has less of its mass concentrated in the extremes, thus less inequality.

Figure 2.1 concludes sections 2.2.1 and 2.2.2 by giving two pedagogical examples.

Figure 2.1: Illustration of X SD$_1$ Y (left) and X MD Y (right).

2.2.3 Latent health approach

Consider the following assumptions about the SRHS variable X, whose categories we enumerate as 1 (poor), 2 (fair), 3 (good), 4 (very good), and 5 (excellent).

Assumption 1. People report health status X according to their "true" latent health X^* and thresholds $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$:

$$X = j \text{ if } \gamma_{j-1} < X^* \leq \gamma_j,\ j = 1, \ldots, 5, \tag{2.2}$$

where $\gamma_0 = -\infty$ and $\gamma_5 = +\infty$.

Assumption 2. Latent health X^* has continuous CDF $F^*(\cdot)$.

Assumption 3 (Stable thresholds)**.** For comparing latent health X^* and Y^*, the corresponding thresholds satisfy $\boldsymbol{\gamma}^X = \boldsymbol{\gamma}^Y + c$, where c is a constant scalar.

Assumptions 1 and 2 are standard in latent variable models. Assumption 3 plays an important part in identifying the latent variance(s) in the inequality context. For example, suppose latent health $X^* \sim \mathrm{N}(\mu_x, \sigma_x^2)$ and $Y^* \sim \mathrm{N}(\mu_y, \sigma_y^2)$. The parameters to estimate are μ_x, σ_x, μ_y, σ_y, γ_1^x, γ_2^x, γ_3^x, γ_4^x, γ_1^y, γ_2^y, γ_3^y, and γ_4^y. These are not all identified given two five-category ordinal variables: the distribution of observables is described by only 10 probabilities ($\mathrm{P}(X = 1)$, $\mathrm{P}(Y = 1)$, etc.), but there are 12 parameters.[6] However, our interest is only in comparing σ_x and σ_y. Assumption 3 implies $\gamma_j^x = \gamma_j^y + c$, where c is a constant. Therefore, X^* and Y^* can be re-scaled as

$$\widetilde{X}^* = \frac{X^* - \gamma_1^x}{\gamma_4^x - \gamma_1^x} = \frac{X^* - \gamma_1^x}{\Delta}, \qquad \widetilde{Y}^* = \frac{Y^* - \gamma_1^y}{\gamma_4^y - \gamma_1^y} = \frac{Y^* - \gamma_1^x + c}{\Delta}.$$

With the rescaling, the parameters reduce to $\{\widetilde{\mu}_x, \widetilde{\sigma}_x, \widetilde{\mu}_y, \widetilde{\sigma}_y, \widetilde{\gamma}_2, \widetilde{\gamma}_3\}$ and are all identified. The relationship ($=$, $>$, or $<$) between $\tilde{\sigma}_x$ and $\tilde{\sigma}_y$ remains the same as that between σ_x and σ_y. Moreover, the term $Y^* - \gamma_1^x + c$ illustrates the argument by Hernández-Quevedo et al. (2005) that, generally speaking, it is not possible to separately identify whether a change in Y's location is due to a shift in the thresholds $\boldsymbol{\gamma}^y$ or due to a shift in the underlying health Y^*.

[6] In other contexts where variance is not a parameter of interest, it is often normalized to equal one. For example, Wagstaff and Van Doorslaer (1994) assume $X^* \sim \mathrm{N}(0, 1)$.

A important question remains: how reasonable is Assumption 3? It is widely agreed that a parallel shift in thresholds exists for age and sex, while it is still unclear if thresholds are further affected differently (i.e., "cut-point shift"). The empirical evidence is mixed. Lindeboom and van Doorslaer (2004) find that the cut-point shift is statistically significant for the young cohort, but not so for the old cohort; Hernández-Quevedo et al. (2005) show little evidence to suggest that reporting bias induced by a change in wording is characterized by a cut-point shift.

Another piece evidence of showing no cut-point shift comes from our own analysis on the PSID. In the year 1999, respondents are asked to describe their health from birth to age 16. Their SRHS before age 16 is distributed in a very similar way for different age cohorts, implying people may have stable thresholds. To sum up, we think Assumption 3 is fairly reasonable.

2.3 Inference

In section 2.3.1, we formally write out the null hypotheses of interest. We show first-order SD and median preserving spread relationships can be written as sets of inequality constraints, which can be jointly tested. Such multiple hypothesis testing has received considerable attention in recent decades. In section 2.3.2, we show how to apply the refined moment selection (RMS) of Andrews and Barwick (2012). In section 2.3.3, we show how to conduct a simple Bayesian inference. It is important to consider both frequentist and Bayesian inference in cases like this since the conclusions may be very different even when the Bayesian prior is noninformative, as explained by Kline (2011) and Kaplan and Zhuo (2017), for example.

2.3.1 Hypotheses

The problem of interest is to assess 1) whether or not a particular inequality ordering exists between two health distributions, and 2) whether their latent variances are equal.

The null hypothesis of latent variance equality is

$$H_0^*\colon \sigma_x = \sigma_y, \quad \text{or equivalently} \quad H_0^*\colon \widetilde{\sigma}_x = \widetilde{\sigma}_y. \tag{2.3}$$

It is straightforward to carry out a popular likelihood ratio (LR) test when latent normality is assumed.

The null hypothesis for first-order SD or for a median preserving spread can be written as a finite number of inequalities. In both cases,

$$H_0\colon \theta_j \leq 0 \text{ for all } j = 1, \ldots, 5. \tag{2.4}$$

The first-order SD null of $H_0\colon F \text{ SD}_1 \, G$ corresponds to (2.4) with $\theta_j = F_j - G_j$, or $\theta_j = G_j - F_j$ to test $H_0\colon G \text{ SD}_1 \, F$. The median preserving spread null of $H_0\colon F \text{ MD } G$ corresponds to (2.4) with $\theta_j = F_j - G_j$ if $j < m$ and $\theta_j = G_j - F_j$ if $m \leq j \leq 5$, where m is the (shared) median category; and vice-versa for $H_0\colon G \text{ MD } F$.

2.3.2 Frequentist testing

When the null hypothesis is "composite" like the H_0 in (2.4), a commonly used device in the frequentist literature is referred to as "the least favorable null." It reduces the composite null hypothesis to a "simple" one by assuming all inequality constraints are binding, i.e., all $\theta_j = 0$ in (2.4). Although this guarantees (asymptotic) size control of a test, it may result in poor power when only a single constraint is violated (and others are satisfied as strict inequalities).

Andrews and Soares (2010) define a testing procedure called generalized moment selection (GMS) to improve power. GMS tries to select only the binding inequalities and then recomputes the worst-case critical value as a function of only these constraints. By not considering the constraints that are easily satisfied, the critical value is smaller, increasing power. In our specific context with SRHS, we can also remove the last inequality constraint from H_0 because $\theta_5 = F_5 - G_5 = 1 - 1 = 0$.

We now describe the refined moment selection (RMS) testing procedure of Andrews and Barwick (2012), building on Andrews and Soares (2010). For testing (2.4), the RMS algorithm can be sketched as follows.

1. For the sample $\{X_i\}_{i=1}^n$, the estimated $\widehat{F}_j$ and estimated asymptotic covariance matrix $\hat{\underline{\boldsymbol{\Sigma}}}(F)$ are

$$\widehat{F}_j = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq j\}; \qquad \widehat{\Sigma}_{jh}(F) = \widehat{F}_b - \widehat{F}_j \widehat{F}_h \tag{2.5}$$

   where $1 \leq j, h \leq 4$ and $b = \min\{j, h\}$. Similarly, estimate $\widehat{G}_j$ and $\hat{\underline{\boldsymbol{\Sigma}}}(G)$ for the sample $\{Y_i\}_{i=1}^m$. The estimated parameter vector of interest is $\hat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \widehat{\theta}_4)$, with $\widehat{\theta}_j = \widehat{F}_j - \widehat{G}_j$. Given the independence between the two samples, the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ is $\hat{\underline{\boldsymbol{\Sigma}}}(\boldsymbol{\theta}) = \hat{\underline{\boldsymbol{\Sigma}}}(F) + \hat{\underline{\boldsymbol{\Sigma}}}(G)$.

2. Compute the adjusted Gaussian quasi-likelihood ratio (AQLR) test statistics for the original sample (see details in Andrews and Barwick (2012)).

3. Simulate R bootstrap samples of sizes n and m (respectively) from the original X and Y samples.

4. In each bootstrap sample, estimate θ_j^b and $\underline{\boldsymbol{\Sigma}}^b(\boldsymbol{\theta})$ by Step 1. Then perform inequality-by-inequality t-tests of the null hypothesis $H_0^j \colon \theta_j \leq 0$ versus $H_1^j \colon \theta_j >$

0 for $j = 1, 2, 3, 4$. The jth inequality constraint is "selected" if

$$\frac{n^{1/2}(\widehat{\theta}_j - 0)}{\widehat{\Sigma}_{jj}(\boldsymbol{\theta})} \leq \widehat{\kappa}, \tag{2.6}$$

where $\widehat{\kappa}$ is a tuning parameter provided by Andrews and Barwick (2012). Then, compute the re-centered AQLR test statistic using the selected constraint(s) and corresponding covariance submatrix from $\hat{\underline{\boldsymbol{\Sigma}}}(\boldsymbol{\theta})$

5. The critical value is the $1 - \alpha$ quantile of the bootstrap distribution of the moment selection version of AQLR test statistic.

2.3.3 Bayesian inference

It is relatively straightforward in the Bayesian framework to make inference on the joint inequality constraints hypothesis H_0 in (2.4). First, Bayesian methods directly provide probabilistic measurement of the constraints, i.e., the posterior probability of H_0. The posterior describes how likely the inequalities hold true given the samples. Second, if desired, posterior probabilities can be converted into a binary decision, such as accept/reject, via a loss function chosen by decision-makers. Moreover, the Bayesian approach can provide coherent, simultaneous inference on all possible relations in the partial ordering. More discussion of Bayesian inferences in the testing context can be found in other studies such as Goutis et al. (1996), DeGroot (2004), and Robert (2007).

Here we briefly introduce the Bayesian Dirichlet–multinomial likelihood model for the ordinal SRHS health variable. Let the number of observations falling in the jth category be denoted by n_j. Thus the data will be represented by $(n_1, \ldots, n_5)$, and their sum is equal to the sample size n. That vector's likelihood function is then multinomial distribution, denoted by $\text{Multi}(p_1, \ldots, p_5;\, n)$. The interpretation is simple: a draw from $\text{Multi}(p_1, \ldots, p_5;\, n)$ can be understood as drawing n iid values from an ordered categorical distribution of X with PMF $f(X = i) = p_i$. We then place a Dirichlet prior, denoted $\text{Dir}(a_1, \ldots, a_5)$, over the population probability vector $\mathbf{p}$. By its conjugacy, the posterior distribution of $\mathbf{p}$ is

$$(p_1, \ldots, p_5) \sim \text{Dir}(a_1 + n_1, \ldots, a_5 + n_5). \tag{2.7}$$

It is easy to sample directly from the posterior in (2.7) and obtain Monte Carlo estimates of various quantities. For example, given a single posterior draw $(p_1, \ldots, p_5)$, the corresponding CDF F_j is $\sum_{i=1}^{j} p_j$.

Our concern here is to summarize the evidence supporting the H_0 in (2.4), in terms of its posterior probability. Lacking any (agreed upon) prior knowledge, one may use a noninformative prior, as below.[7] Under the assumption of independent X and Y samples, the posterior of H_0 is computed by the following algorithm.

1. The posterior for X is $\mathbf{p} \equiv (p_1, \ldots, p_5) \sim \text{Dir}(n_1 + 1, \ldots, n_5 + 1)$. Similarly, the posterior for Y is $\mathbf{q} \equiv (q_1, \ldots, q_5) \sim \text{Dir}(m_1 + 1, \ldots, m_5 + 1)$.

2. Draw R posterior samples of $\mathbf{p}$ and $\mathbf{q}$. In each, compute the corresponding $\boldsymbol{\theta}$. For the rth draw, let $I_r = 1$ if $\boldsymbol{\theta}$ satisfies H_0, otherwise $I_r = 0$.

3. The (approximated) posterior probability of H_0 is $R^{-1} \sum_{r=1}^{R} I_r$.

2.4 Examples

The data used here are from the Current Population Survey (CPS). For the analysis presented here, the Annual Social and Economic (ASEC) supplement to CPS is used. It has collected information about health and benefits since 1994. The ASEC is a repeated cross-sectional survey and contains SRHS data on different cohorts.

[7] There are multiple "noninformative priors" in this case, but the differences are practically negligible with even moderate sample sizes, since they all have $0 \leq a_j \leq 1$ for each j.

2.4.1 Dynamics of health over the life cycle

The first example is concerned with how dispersion of health evolves with age, which is the question raised by Deaton and Paxson (1998). Their hypothesis is that health inequality increases with age. But we will show that evidence from SRHS alone may not be adequate for supporting this hypothesis.

Table 2.1 gives an overview of a group of individuals born between the years 1972 and 1976. Thanks to the richness of the data, we can follow this same birth cohort every five years and have up to five cohort-year pairs, which are non-overlapping. For example, respondents in this cohort would age from 20 to 24 years old in year 1996, and from 25–29 in 2001. It is clear that the SRHS distribution is deteriorating with age. Also, the demographic composition is relatively stable.

Table 2.1: Summary statistics for birth cohort born between 1972 and 1976

Age	Wave year	Obs.	Mean	SD	Black (%)	Male (%)
[20, 24]	1996	8093	4.13	0.89	11.19	47.65
[25, 29]	2001	13 126	4.07	0.92	12.36	47.46
[30, 34]	2006	13 589	4.01	0.95	10.62	47.52
[35, 39]	2011	13 928	3.91	0.97	10.35	47.60
[40, 44]	2016	12 381	3.82	1.01	11.79	48.39

Excellent = 5, Very good = 4, Good = 3, Fair = 2, Poor =1

Figure 2.2 presents the proportion of each health status category at five different stages of life. From the top panel, it shows a steady increase in the cumulative fraction at every level of health for both males and females. Regarding race, it is still true for whites. But blacks seem to become healthier when moving into the second half of their 20s, after which they have similar declining trend.

In table 2.2, we conduct pairwise tests among five black cohort-year groups. When testing the equality of latent variances, the LR test only rejects the null twice when comparing the youngest with the oldest and the second-oldest. It implies the latent health variance does change over time, but not as dramatically as the ordinal "vari-

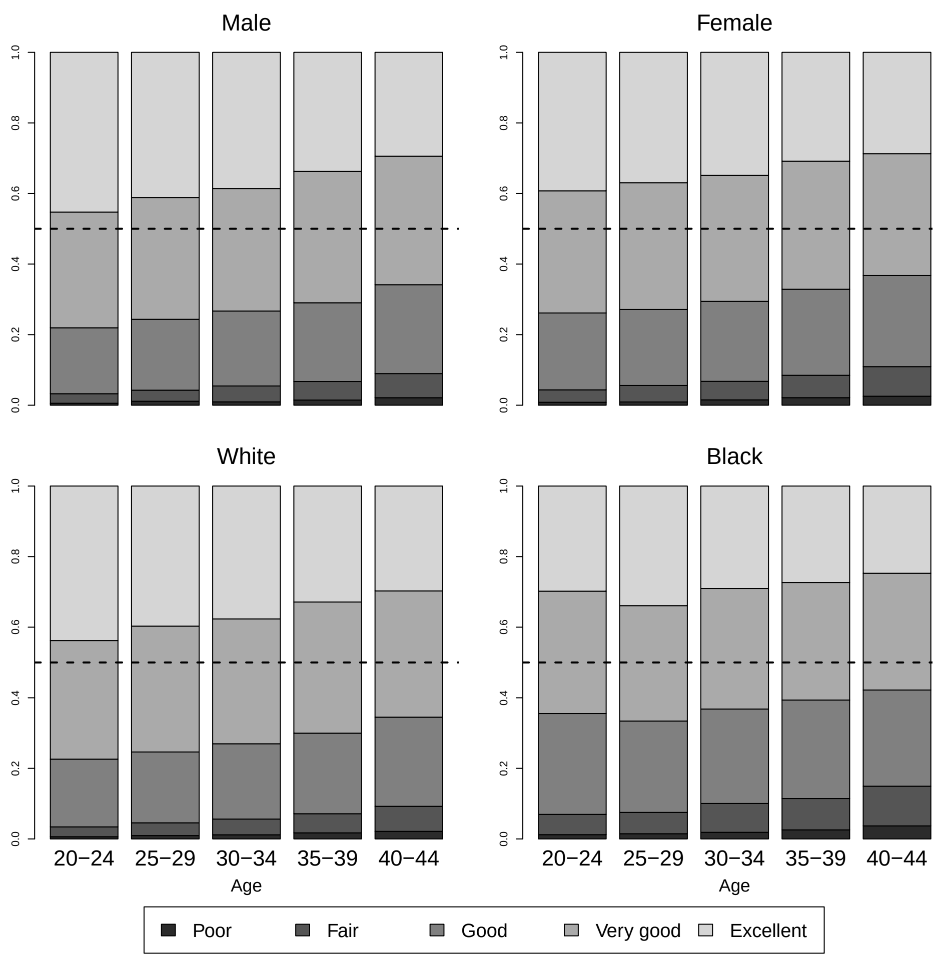


Figure 2.2: Fraction of heath status category by sex (top row) or race (bottom).

ances" suggest. On the other hand, the change in the latent variance is not fully revealed in ordinal partial orderings since the group of younger individuals almost always first-order stochastically dominates the older group. That is, the declining health is so overwhelming that it swamps information about health inequality. The argument we want to make is, when first-order SD is present in the ordinal data, it may be difficult or infeasible to elicit answers to questions of inequality or similar subjects.

Table 2.2: Inference for health changes with age, for black cohort born 1972–1976.

		$H_0^*: \sigma_X = \sigma_Y$	$H_0: X \text{ SD}_1\ Y$		$H_0: Y \text{ SD}_1\ X$	
X	Y	LR	RMS	Bayes	RMS	Bayes
[20, 24]	[25, 29]	6.35 %	8.0 %	**0.8 %**	63.6 %	13.1 %
	[30, 34]	8.90 %	100 %	42.3 %	**3.7 %**	**0.1 %**
	[35, 39]	**2.68 %**	100 %	74.7 %	**0.1 %**	**0.0 %**
	[40, 44]	**0.20 %**	100 %	86.0 %	**0.0 %**	**0.0 %**
[25, 29]	[30, 34]	85.63 %	100 %	67.2 %	**0.0 %**	**0.0 %**
	[35, 39]	66.67 %	100 %	85.2 %	**0.0 %**	**0.0 %**
	[40, 44]	13.94 %	100 %	86.0 %	**0.0 %**	**0.0 %**
[30, 34]	[35, 39]	54.21 %	100 %	57.0 %	**1.8 %**	**0.5 %**
	[40, 44]	9.54 %	100 %	88.3 %	**0.0 %**	**0.0 %**
[35, 39]	[40, 44]	28.95 %	100 %	75.7 %	**1.0 %**	**0.0 %**

1. p-value for LR and RMS; posterior probablity of H_0 for Bayes.
2. Entries in **bold** indicate rejection at 5% significance level.

Though the Bayesian and frequentst (RMS) procedures agree most of time, they conclude differently for the comparison between the age ranges $X = [20, 24]$ and $Y = [25, 29]$. There, RMS fails to reject SD1 (p-value of 8.0%), but the Bayesian posterior probability of X SD_1 Y is only 0.8%. Given this disrepency, we further study the second partial ordering, median preserving spread.

Table 2.3 is similar to table 2.2 for age ranges $X = [20, 24]$ and $Y = [25, 29]$, but for median preserving spread instead of SD1. RMS again fails to reject all possible partial orderings, with the highest p-value for Y SD_1 X. The Bayesian test rejects the nulls of both X SD_1 Y and Y MD X (i.e., posterior probability is below 5%).

Table 2.3: Testing partial orderings between age ranges $X = [20,\ 24]$ and $Y = [24,\ 29]$, 1972–1976 black cohort.

	$H_0: X\ \mathrm{SD}_1\ Y$	$H_0: Y\ \mathrm{SD}_1\ X$	$H_0: X\ \mathrm{MD}\ Y$	$H_0: Y\ \mathrm{MD}\ X$
p-value	8.0 %	63.2 %	37.2 %	9.3 %
Post. prob.	**0.8 %**	13.1 %	7.7 %	**0.0 %**

1. Entries in **bold** indicate rejection at 5% significance level.

2.4.2 Trends in health inequalities for the elderly cohorts

Older Americans are living longer and health expenditure is booming. But are they living healthier? Is population health today more equal than before? In this section, SRHS data are used to answer these two questions, using the methodology from section 2.3.

To study the inter-cohort trends in health inequality for the elderly, we focus on five cohorts born in different periods and select their SRHS data when aged between 65 and 70.[8] Table 2.4 provides descriptive statistics for the cohorts. The "average" health rating is getting better and the "standard deviation" is decreasing over generations.

Table 2.4: Sample: cohorts of the elderly aged 65-70

Cohorts	Birth year	Wave year	Obs.	Mean	SD	Black (%)	Male (%)
SG_1	1926-1931	1996	5642	3.03	1.17	8.06	46.44
SG_2	1932-1936	2001	7203	3.09	1.13	13.66	46.19
SG_3	1937-1941	2006	7160	3.13	1.13	12.12	46.98
SG_4	1942-1946	2011	8145	3.17	1.11	12.62	45.84
EBB	1947-1951	2016	9726	3.20	1.11	13.26	47.12

1. Excellent = 5, Very good = 4, Good = 3, Fair = 2, Poor = 1.
2. SG, or the Silent Generation, includes 4 cohorts, from early SG_1 to late SG_4; cohort EBB is early baby boomers.

Figure 2.3 presents the proportion of each health status category for the five cohorts. The top panel show that a steady health improvement has been seen over generations, regardless of race. When broken down by sex and race, the black male

[8] We purposely choose this age interval to limit the effect of survivorship bias (Heiss, 2011; Heiss et al., 2014) and the effect of financial burden (e.g., full retirement age starts at 65).

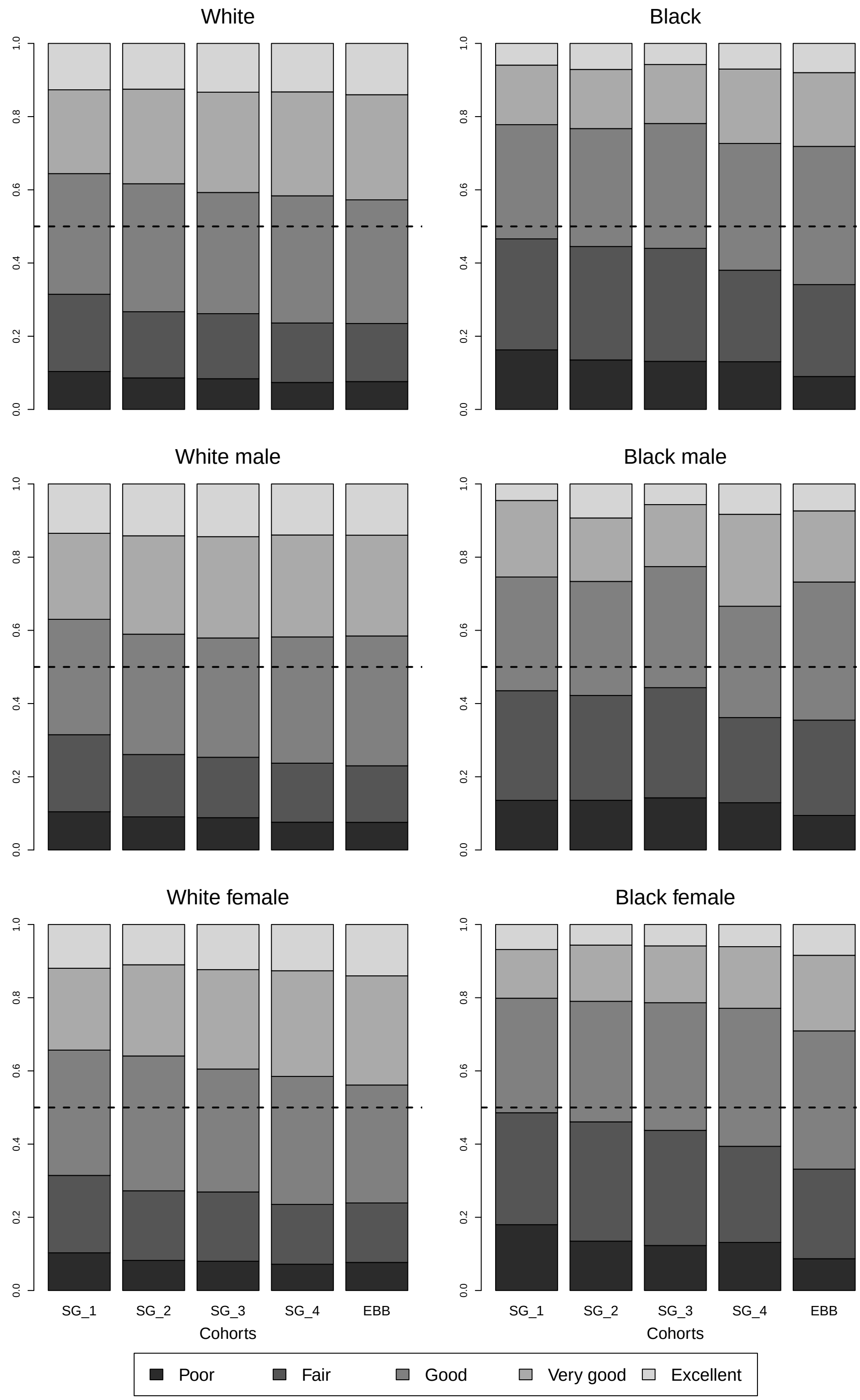


Figure 2.3: Fraction of SRHS category for the elderly by sex or race

group shows different patterns of changes in the fraction at each health category. Therefore, our tests are carried out with emphasize on black males. In particular, we investigate the partial orderings between SG_3 and SG_2 for black individuals (all, male, female), and those between SG_4 and EBB for males (all, black, white). In table 2.5, results from both frequentist and Bayesian methods are provided. Bayesian interpretation will be used for the sake of probabilistic measurement.

Table 2.5: Comparing different birth cohorts at ages 65–70

			$H_0\colon X$ MD Y		$H_0\colon X$ SD_1 Y		$H_0\colon Y$ SD_1 X	
X	Y	Sample	RMS	Bayes	RMS	Bayes	RMS	Bayes
SG_3	SG_2	black (all)	100 %	25.8 %	39.7 %	**3.2 %**	100 %	17.9 %
		black male	76.4 %	11.7 %	12.4 %	**0.1 %**	100 %	39.4 %
		black female	100 %	10.3 %	100 %	25.4 %	61.2 %	**4.3 %**
EBB	SG_4	male (all)	100 %	26.9 %	51.4 %	9.3 %	58.9 %	6.8 %
		black male	100 %	38.3 %	**4.6 %**	**1.2 %**	16.4 %	**2.0 %**
		white male	100 %	12.4 %	100 %	15.8 %	65.3 %	10.3 %

1. p-value for LR and RMS; posterior probablity of H_0 for Bayes.
2. Entries in **bold** indicate rejection at 5% significance level.

Table 2.5 shows results from frequentist (RMS) and Bayesian inference on both SD1 and median preserving spread relationships, for certain demographic groups and cohorts. In terms of median preserving spread, the Bayesian posterior probability for black individuals in SG_3 being more equal than those in SG_2 is 25.8%, but it is lower for both black males and black females. This suggests that the improvement in inequality may result from the closing gap between black males and females. This is supported by results for the black male and black female groups: for black males, there is more evidence that SG_3 first-order stochastically dominates SG_2 (i.e., SG_3 is less healthy), whereas it is the opposite for black females. For the second comparison, the Bayesian posterior assigns 26.9% probability of EBB males being more equal than ones in SG_4, in the sense of median preserving spread. For black males, the Bayesian test rejects both SD hypotheses at a 5% level, and it assigns 38.3% posterior probability to the EBB cohort having less inequality.

2.5 Conclusion

Neither the parametric latent variable approach nor the median preserving spread is fully satisfactory in measuring inequality based on ordinal data. The latent variable approach is not robust to misspecification, while the requirement of common medians for the median preserving spread is not always satisfied. Nonetheless, the simple SRHS variable can be exploited to compare health distributions.

This paper attempts to show what the ordinal variable is capable of answering in two different contexts. When studying the age effect on health, we find SRHS is able to answer how health changes with age, but it fails to account for the dynamics of health inequality. For the cohort effect on health inequality, we take advantage of the common median and redefine the inequality concept.

Empirical researchers often choose the best available measures of quantities and concepts of interest. Nevertheless, it is important to bear in mind there always are boundaries of what questions the best available measures can answer. This paper provides evidence on both the limitations and abilities of SRHS to provide insight into health inequality.

Chapter 3

Portfolio selection using stochastic dominance and nonparametric Bayesian method

3.1 Introduction

Stochastic dominance (SD) is a well-established rule for investment decision making under uncertainty. Its superiority comes from the fact that it avoids the usual normal approximation to the return distributions and, more importantly, imposes few restrictions on specification of investor preferences (Hadar and Russell, 1969; Hanoch and Levy, 1969; Rothschild and Stiglitz, 1971; Bawa, 1975; Levy, 2015). This ordering rule is particularly appealing for investment strategy and asset classes with higher-order moment risk, like small-cap stocks and junk bonds. In these cases, the traditional variance does not fully capture the asymmetric risk profile, as it fails to distinguish between upside potential ("good risk") and downside risk ("bad risk").

Portfolio construction based on SD is a theoretically appealing alternative to Markowitz's Mean-Variance (MV) analysis, given its robustness on return distributions and welfare/utility implication for a broad class of investors. A popular criterion is the second-order SD (SSD), which can be defined by conditional value at risk (CVaR), a widely used risk measure (Dentcheva and Ruszczyński, 2003). Besides, any risk-averse investor would strictly prefer the portfolio enhanced by SSD criterion. However, its applications face the following difficulties. First, SSD-enhanced portfolio should be considered from all possible portfolios, each of which compares with the benchmark; that is, the problem involves infinitely many pair-wise SD comparisons. Second, a distribution-free assumption in SD requires the nonparametric statistical inference methods on return distributions.

The operations research (OR) literature mainly focuses on solving the first issue. The typical optimization problem with SSD constraints often assumes that a reference portfolio (i.e., benchmark) is available and another portfolio is constructed, whose return distribution dominates the benchmark with respect to SSD. In most cases the problem has a large number of constraints, since it involves a infinitely large number of pairwise SD relations and each relation involves a large number of inequalities. Dentcheva and Ruszczyński (2003, 2006) consider the problem of constructing such optimal portfolio of finitely many assets whose return is discretely distributed. The discreteness assumption enables them to develop, based on the theory of majorization, a linear programming (LP) where the objective is to maximize the portfolio expected return with SSD constraints over the benchmark. However, the size of LP problem grows at a quadratic rate with the number of observations and becomes very large in applications with hundreds or thousands of possible outcomes. To avoid this obstacle, Luedtke (2008) describes a compact linear programming formulation based on the Strassen Theorem, which greatly reduces the number of constraints needed for SSD requirements. Alternatively, Roman et al. (2006) propose a multi-objective optimization problem whose Pareto optimal solution portfolio second-order dominates the benchmark. A particular solution is chosen whose return distribution comes close to the benchmark in a uniform sense. Uniformity is defined by the differences among tails risk (or, CVaR). In practice, it does not work well since the tail risk is treated equally in the optimization. Therefore, Fábián et al. (2011) describe an enhanced version of the multi-objective model, which compares the scaled values of different objectives. Such scaled objectives reflect different confidence levels at the tail risks of a return distribution. An efficient algorithm, called the cutting-plane representation, is applied by Fábián et al. (2011) for the SSD multi-objective optimization. Roman et al. (2013) investigate it through re-balancing and back-testing by using several data sets, including SP 500 and FTSE 100.

Though the OR literature successfully address the first issue, most of them ignore another: the statistical inference of the joint return distribution of base assets. The statistical accuracy is particularly important in portfolio optimization, because the optimal portfolio weights can be very sensitive to estimation error and then the constructed portfolios may have a quite poor out-of-sample performance.

The econometrics and finance literature mainly focuses on testing hypotheses of dominance or non-dominance for a given set of choice alternatives. Among others, Davidson and Duclos (2000) hypothesize SD between two distributions as inequality constraints at a fixed number of arbitrary chosen points and derive the asymptotic sampling distribution of related test statistics. Barrett and Donald (2003) advance such test of pair-wise SD relation by checking the inequalities at all points in the support of the distribution. Linton et al. (2005) go beyond the pair-wise dominance and propose a general test for the general K random variables (or, distributions). On the other hand, the test of *non*SD null is advocated due to its analytical convenience (Kaur et al., 1994) and practical usefulness (Davidson and Duclos, 2013). In fact, the null of *non*SD is widely used in finance. Post (2003) defines a portfolio is SSD efficient if and only if it is not dominated by any other feasible portfolio. He further develops an easy-to-implement LP test for the SD efficiency of a particular portfolio relative to all possible portfolios constructed from base assets. Linton et al. (2014) further improve the power of the LP-type stochastic dominance efficiency test. Nevertheless, the definition of SD efficiency is not exclusive. For example, Kuosmanen (2004) defines a portfolio is SD efficient if it dominates all alternative portfolios. Scaillet and Topaloglou (2010) examine the version of SD efficiency in Kuosmanen (2004) and develop a Kolmogorov–Smirnov (KS) type test based on the Barrett and Donald (2003)'s approach.

These studies unfortunately offer little guidance for constructing a dominant portfolio with full diversification possibilities, though they certainly provide a stimulus to the further study for research relevant to portfolio selection and evaluation. The existing optimization problem with SSD constraints generally uses the empirical distribution function (EDF). This approach is statistically accurate when the available time series is long and the number of base assets is relatively moderate. In practice, however, the sample size is relatively small and the uncertainty about population distribution function becomes the problem encountered by all investors.

This study first evaluates current optimization formulations using same data set. It is meaningful in terms of comparing their validity and performances. Second, we proposes a new formulation in order to account for statistical uncertainties. These estimation errors are quantified by the probabilistic constraints and nonparametric Bayesian (NPB) inference (Ferguson, 1973, 1974; Rubin, 1981), such that the optimal solution should satisfy the SSD relation with a prescribed (posterior) probability.

Bayesian inference has several desirable statistical properties and information-theoretic implications. In the literature of portfolio selection and analysis, model uncertainty and parameter uncertainty are two important problems. Avramov and Zhou (2010) advocate the use of Bayesian framework for these uncertainties and review many Bayesian portfolio studies, including the seminal work of Zellner and Chetty (1965), prior update (Black and Litterman, 1992), asset pricing prior (Pástor, 2000; Pástor and Stambaugh, 2000). Though the informative / subjective prior is extensively used in these researches, the focus of our NPB method is on noninformative asset prior

We consider the realized return vectors as the support of a multinomial distribution.[1] The posterior probability for each return vector is computed by the NPB model of Ferguson (1973, 1974). Specifically, we use the approach of Rubin (1981) to implement with an improper and noninformative Dirichlet process prior. This approach preserves the information of the historical returns and their cross-sectional dependence. Meanwhile, it allows for a finite, state-dependent representation of the portfolio optimization problem. The probabilistic feature of Bayesian method can naturally used to construct the stochastic programming with probabilistic SSD constraints.

Importantly, NPB combines well with SD, due to a shared nonparametric assumption on distribution function. The complementary relation between NPB and SD was recognized earlier by Zhuo (2017) and Kaplan and Zhuo (2017). Those earlier studies use NPB to test stochastic dominance relations among two random variables/distributions. By contrast, our study uses the posterior distribution of population distribution for SSD-enhanced portfolio's construction. The fusion of SD and NPB better serves to account for model uncertainty and parameter uncertainty by a tractable mixed-integer linear programming.

Post and Karabati (2016) conduct a study close to ours. They develop a portfolio optimization method based on SSD and the empirical likelihood (EL) estimation method. Their SD/EL method can be implemented using by two steps. Step one is to elicit the EL probabilities by minimizing the Kullback–Leibler (KL) divergence; Step two is to plug these probabilities into LP for the optimal portfolio construction.

We investigate SSD optimization formulations using two types of data sets. One data set is weekly return to 442 equities (stocks) from 11/2004 to 04/2016, 592 weeks

[1] For each cross section, returns of base assets is a vector. This vector is one of many points in the support for the joint return distribution.

in total. The benchmark is the Standard & Poor's 500 (SP 500). Another one is daily returns to 49 equity industry portfolios from 01/03/2000 to 12/30/2016. It should be noted that base assets are portfolios, upon which the optimal portfolio is built. We use the heuristic weighted portfolio as the benchmarks, and also consider the mean-variance weights for comparisons. Since the objective is active portfolio selection, we re-balance the optimal portfolio periodically and do not assume that the benchmark portfolio is efficient.

The rest of paper is organized as following: Section 3.2 discusses the formulation of SSD-based optimization, nonparametric Bayesian methods and a new formulation; Section 3.3 apply the methods to two empirical examples. All conclusions and suggestions are in the Section 3.4.

3.2 Methodology

This section presents two major SSD-efficient optimization formulations and proposes a new formulation based on Mixed Integer Linear Programming (MILP) and posterior probability of NPB.

3.2.1 Preliminaries

Suppose we have n distinct base assets with random return vector $\mathbf{r} = (r_1, \ldots, r_n) \sim F(\cdot)$, where $F(\cdot)$ is the *joint* Cumulative Distribution Function (CDF) with the support of $\mathcal{R} \subset \mathbb{R}^n$.[2] We can define any portfolio X by a convex combination of base assets as follows:

$$X(\boldsymbol{\lambda}) = r_1 \lambda_1 + \ldots + r_n \lambda_n = \mathbf{r}^\intercal \boldsymbol{\lambda}$$

[2]It should be noted that the base assets are not limited to individual securities/equities. For example, they can be portfolios.

where the asset weights $(\lambda_1, \lambda_2, \ldots, \lambda_n) \in \Lambda \equiv \left\{\boldsymbol{\lambda} \in \mathbb{R}^n : \boldsymbol{\lambda}^\intercal \mathbf{1}_n \ = 1,\ \boldsymbol{\lambda} \geq \mathbf{0}_n\right\}$. Here we only consider portfolio construction based on the long strategy, though the sell short is also important.

The marginal CDF of portfolio X is given by

$$F_{\mathbf{X}}(z) \equiv \int_{\{\mathbf{r}\in\mathcal{R}:\ \mathbf{r}^\intercal\mathbf{x}\leq z\}} dF(\mathbf{r}) \tag{3.1}$$

Let Y be random return of a particular reference/benchmark portfolio of interest. We attempt to find a set of portfolio weights $\boldsymbol{\lambda}$ such that its portfolio return, i.e., $X = \mathbf{r}^\intercal\boldsymbol{\lambda}$, second-order stochastically dominates Y.

Definition 3. A portfolio X second-order stochastic dominates (SSD) the benchmark Y, or X SD_2 Y if and only if any one of the following equivalent conditions is satisfied:

1. $\displaystyle\int_{-\infty}^{z} F_X(v)dv \leq \int_{-\infty}^{z} F_Y(v)dv,\ \ \forall z \in \mathcal{R}$
2. $\displaystyle\int_{0}^{\alpha} Q_X(v)dv \geq \int_{0}^{\alpha} Q_Y(v)dv,\ \ \forall \alpha \in (0,1)$, where $Q(\cdot)$ is the quantile function
3. $\mathbb{E}_u(X) \geq \mathbb{E}_u(Y),\ \ \forall u \in \mathcal{U} \equiv \{u(\cdot) : u' \geq 0, u'' \leq 0\}$

In the first and second definitions of the SSD criterion, random portfolio return rates are compared by the distribution function or quantile funciton. There are fundamental relations between these functions and risk management. For example, the integral of the distribution function is connected with the expected shortfall in the finance by

$$\int_{-\infty}^{z} F_X(v)dv \leq \int_{-\infty}^{z} F_Y(v)dv \iff \mathbb{E}\big[(z-X)_+\big] \leq \mathbb{E}\big[(z-Y)_+\big] \tag{3.2}$$

where the shortfall $(z-X)_+ = \max(z-X, 0)$. Then X SD_2 Y iff the expected shortfall of X is less for X at any level Z. Furthermore, The connection between the quantile function and Conditional Value at Risk (CVaR) can be established as, for $\forall \alpha \in (0, 1)$

$$\mathrm{CVaR}_{\alpha}(X) = -\int_{0}^{\alpha} Q_X(v)dv.$$

Beyond it, the economic interpretation of SSD is shown in the third definition, which suggests, for any increasing and concave utility function, the expected utility under the return distribution X would be higher and then strictly preferred. In short, SSD has strong connection with important risk measures such as CVaR, and strong implication in the welfare theory.

The original concept of stochastic dominance can apply to only two random variables / distributions. It is clearly not applicable when it comes to an infinite set of diversification strategies used in the portfolio construction. To overcome it, Post (2003) and Kuosmanen (2004) extend the pair-wise stochastic dominance relation and discuss one type of general dominance relations among infinitely many alternatives, called stochastic dominance (in)efficiency.

Definition 4. A portfolio Y is SSD *in*efficient if and only if there exists some portfolio X such that X SD_2 Y. Alternatively, portfolio Y is SSD efficient if and only if no portfolio X SD_2 Y.

In this study, we assume the benchmark portfolio Y is SSD inefficient.[3] Our goal is to find a set of portfolio weights ($\boldsymbol{\lambda}$), such that the portfolio X $= \mathbf{r}^{\intercal}\boldsymbol{\lambda}$ can dominate Y at the second order. It is done by a portfolio optimization with SSD constraints. In general, these constraints are not trivial from many aspects: the constraints are not linear and the number of these constraints is infinitely large. However, as we will see later, the discreteness return assumption allows us to build a finite system of linear constraints for this problem.

[3]Many statistical tests are proposed to study the stochatic dominance efficiency (Post, 2003; Kuosmanen, 2004; Scaillet and Topaloglou, 2010; Linton et al., 2014).

It is worthy discussing why SSD efficiency is better than mean-variance (MV) efficiency. The MV paradigm is valid in the sense of von Neumann–Morgenstern utility maximization axioms if one of the following assumptions are true: 1. the utility function is quadratic (e.g., second-degree polynomial); 2. the portfolio return distribution is from a two-parameter exponential family and the utility function is concave (Baron, 1977). If both conditions are satisfied, the MV approach is the simplest and best. But it is often not the case. Porter and Gaumnitz (1972) claim the SD rule is less restrictive than MV rule. They further claim that highly risk-averse investors may violate the MV maximization of expected utility, while low risk-averse investors are indifferent to MV and SD rule. In short, SD rule is more comprehensive and robust approach to manage risk than MV rule.

3.2.2 Formulation of SSD optimization

The SSD optimization can be understood as an enhanced indexing strategy in finance. It attempts to outperform the benchmark indexing/portfolio Y by finding a new set of portfolio weights $\boldsymbol{\lambda}$. In particular, the SSD rule enhances the benchmark Y by managing its tail risk, that is, reduce the downside (bad) risk and increase the upside (good) risk.

A generic SSD optimization problem, under some regular assumptions, can be summarized as

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & f(\boldsymbol{\lambda}) \\ \text{subject to} \quad & X(\boldsymbol{\lambda}) \quad \mathrm{SD}_2 \quad Y \\ & \boldsymbol{\lambda} \in \Lambda. \end{aligned} \tag{3.3}$$

It is a stochastic optimization in the sense that X and Y are in a probability space $(\Omega, \mathcal{F}, P)$. Generally, there are two lines of formulations, depending on how $f(\cdot)$ is specified. Dentcheva and Ruszczyński (2003, 2006) pioneer the first kind of formulation, whose objective is to maximize the expected return of a portfolio subject to the second-order stochastic dominance. The second approach is proposed by Roman et al. (2006). Their objectives are, instead of expected return, the differences of CVaR_α between the benchmark and the portfolio by $\boldsymbol{\lambda}$, for any $\alpha \in (0, 1)$. In other words, their approach is a multi-objective framework with emphasis on risk management.

The weights $\boldsymbol{\lambda}$ must satisfy the stochastic constraints $X(\boldsymbol{\lambda}) \quad \mathrm{SD}_2 \quad Y$. If X and Y are continuous, it implies the number of inequalities is infinitely many, from which the issue of feasibility may arise. To get around it, the discreteness of the joint CDF $F(\cdot)$ for base assets is assumed. Therefore, X and Y are also discrete.

Some notation in this section will be described in Table 3.1.

Table 3.1: Notation for SSD optimization problem

Symbol	**Description**
n	Number of assets
T	Number of periods/observations
λ_i	Portfolio weight for asset i
r_{it}	Rate of return (RoR) for asset i at time t
X	RoR for the enhanced portfolio $\mathbf{r}^\intercal\boldsymbol{\lambda}$
Y	RoR for the benchmark portfolio Y,
	Y is a discrete variable with finite supports
D	Number of finite supports for Y
x_t	RoR of X at time t
p_t	Probability for x_t, or for r_{it}, $\forall i$
y_k	Ordered RoR of Y, (i.e., $y_1 \leq \cdots \leq y_k \leq \cdots \leq y_D$)
q_k	Probability for $Y = y_k$

Return-based SSD optimization formulation

Dentcheva and Ruszczyński (2003) are interested at the expected return of enhanced portfolio, that is

$$f(\boldsymbol{\lambda}) = \mathbb{E}[X(\boldsymbol{\lambda})] = \mathbb{E}[\mathbf{r}^\intercal\boldsymbol{\lambda}].$$

They seek the highest expected return of portfolios among those dominating the benchmark w.r.t. SSD. This specific objective function is inspired by the MV framework, which is to maximize the expected return under the constraint of variance. Here, the variance-based constraint in MV is replaced by a more robust SSD decision criterion, though the problem also becomes more complicated. Such trade-off is worthwhile: the asymmetric risk profile is a common feature in asset markets. Variance fails to recognize it and treats the upward risk as well as downward risk, while SSD decision rule can effectively distinguish them.

They also show that the SSD constraints in Equation (3.3) define a convex feasible region, regardless what kind of objective function is used. Further, they introduce a decision variable $s = z - \mathbf{r}^\intercal\boldsymbol{\lambda}$ for solving the non-smoothness in $\mathbb{E}\big[(z - \mathbf{r}^\intercal\boldsymbol{\lambda})_+\big]$.

A return-based version of Equation (3.3) is

$$\begin{aligned}
\max_{\boldsymbol{\lambda},\, s_{kt}} \quad & \mathbb{E}[\mathbf{r}^\intercal\boldsymbol{\lambda}] \\
\text{s.t.} \quad \sum_{i=1}^{n} \lambda_i r_{it} + s_{kt} &\geq y_k && k = 1, \dots, D,\ t = 1, \dots, T \\
s_{kt} &\geq 0 && k = 1, \dots, D,\ t = 1, \dots, T \\
\sum_{t=1}^{T} p_t s_{kt} &\leq F_2(Y; y_k), && k = 1, \dots, D \\
\lambda_i &\geq 0 && i = 1, \dots, n \\
\sum_{i=1}^{n} \lambda_i &= 1
\end{aligned} \tag{3.4}$$

where $F_2(Y; y_k) = \mathbb{E}\big[(y_k - Y)_+\big] = \sum_{j=1}^{D} q_j (y_k - y_j)_+$.

We refer to this formulation as *DR*. It is a linear programming with $n + DT$ variables (i.e., λ and s) and $(D+1)T$ constraints.[4] The optimal portfolio $X = \mathbf{r}^\intercal\boldsymbol{\lambda}^\star$ to this system dominates the benchmark portfolio Y by SSD. Correspondingly, we have the optimal solution $s_{kt}^\star = \max\big\{0, z_k - X_t\big\}$.

[4]Many linear programming solvers, such as free open-source R package "lpSolve" and commercial solver IBM Ilog CPLEX, are readily available for solving it.

Risk-based SSD optimization formulation

Roman et al. (2006) explore the connections of SSD with Conditional Value at Risk (CVaR) at different levels. The CVaR of a random return R at $\alpha \in [0,1]$ can be understood as a way to describe the expected losses in the worst $\alpha \times 100\%$ of scenarios.[5] Fábián et al. (2011) define another relevant concept, $\mathrm{Tail}_\alpha(R)$, which is the unconditional expectation of the least $\alpha \times 100\%$ outcomes of the random variable R. That is

$$\mathrm{Tail}_\alpha(R) = \int_0^\alpha Q_R(v)dv = -\mathrm{CVaR}_\alpha(R) \tag{3.5}$$

Then the SSD above can be written as

$$X \quad \mathrm{SD}_2 \quad Y \qquad \Longleftrightarrow \qquad \mathrm{Tail}_\alpha(X) \;\geq\; \mathrm{Tail}_\alpha(Y),\ \forall \alpha \in (0,1)$$

In the case of discrete return with equal probability for all D support points, they propose a multi-objetive approach, in which the D objective functions can be written as Tail at D different levels. Therefore, the SSD efficient portfolios are Pareto optimal solutions to the following multi-objective model:

$$f(\boldsymbol{\lambda}) = \Big(\mathrm{Tail}_{\frac{1}{D}}(\mathbf{r}^\intercal\lambda), \ldots, \mathrm{Tail}_{\frac{i}{D}}(\mathbf{r}^\intercal\lambda), \ldots, \mathrm{Tail}_{\frac{D}{D}}(\mathbf{r}^\intercal\lambda)\Big) \tag{3.6}$$

where $f(\cdot)$ is a $D \times 1$ vector and $\boldsymbol{\lambda} \in \Lambda$.

Moreover, the reference-point method is needed in order to choose a particular SSD efficient solution, whose return distribution comes closest, in a uniform sense, to the benchmark portfolio Y. It can be done by a single objective optimization problem. Let a reference point be

$$\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \ldots, \widehat{\tau}_D) \equiv \Big(\mathrm{Tail}_{\frac{1}{D}}(Y), \ldots, \mathrm{Tail}_{\frac{D}{D}}(Y)\Big).$$

[5]The formal definition of CVaR is given for example in Rockafellar and Uryasev (2000, 2002).

The reference point method also introduces a concave "achievement function" for each element of the objective in Equation (3.6). For example, the simplest achievement function is

$$\Gamma_{\boldsymbol{\tau}}(\tau_1, \ldots, \tau_D) \equiv \min_{1 \le i \le D} (\tau_i - \widehat{\tau}_i)$$

Different achievement functions yield different results, which can reflect the preference of the model developer. In particular, under the achievement function above, the single-objective optimization problem takes the form

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \Gamma_{\widehat{\boldsymbol{\tau}}}\Big(\mathrm{Tail}_{\frac{1}{D}}(\mathbf{r}^{\mathsf{T}}\lambda), \ldots, \mathrm{Tail}_{\frac{D}{D}}(\mathbf{r}^{\mathsf{T}}\lambda)\Big) \\ \text{subject to} \quad & \boldsymbol{\lambda} \in \Lambda. \end{aligned}$$

Letting $\vartheta = \min_{1 \le i \le S} \Big(\mathrm{Tail}_{\frac{i}{S}}(\lambda^{\mathsf{T}}\mathbf{r}) - \widehat{\tau}_i\Big)$, the worst partial achievement, we can re-write the problem above as

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & \vartheta \\ \text{s.t.} \quad & \vartheta \in \mathbb{R}, \quad \lambda \in \Lambda \\ & \vartheta \le \mathrm{Tail}_{\frac{i}{S}}(\mathbf{r}^{\mathsf{T}}\boldsymbol{\lambda}) - \widehat{\tau}_i \qquad \forall i = 1, \ldots, D. \end{aligned} \tag{3.7}$$

Many new constraints are introduced, due to the fact that $\boldsymbol{\lambda}$ is an optimizer. In fact the number of new constraints is on the order of D^2. Therefore, Fábián et al. (2011) proposed a cutting-plane approach to compute the quantities $\mathrm{Tail}_{\frac{i}{S}}(\mathbf{r}^{\mathsf{T}}\boldsymbol{\lambda})$, which proves faster and more effective on larger number of constraints. The idea is to create a sequence of iterates $\boldsymbol{\lambda}_{\mathbf{1}}, \boldsymbol{\lambda}_{\mathbf{2}}, \ldots \in \Lambda$; at each iterate $\boldsymbol{\lambda}$, we check each constraint one by one, if the current iterate becomes a feasible solution of the original problem, stop moving to the next iterate; otherwise, the violated constraint would define a plane, which cuts the Λ into two halves. We move to the next iterate and make sure the new $\boldsymbol{\lambda}$ stays in the half where the constraint violated before is satisfied this time; keep searching until every constraint is satisfied. Surprisingly, it usually needs only a small number of cutting-planes before reaching the optimal solution.

Fábián et al. (2011) propose a scaled version of RMZ, which uses the "scaled" $\text{Tail}_{\frac{i}{S}}(\mathbf{r}^\intercal\boldsymbol{\lambda})$ as follows:

$$
\begin{aligned}
\max_{\boldsymbol{\lambda}} \quad & \vartheta \\
\text{s.t.} \quad & \vartheta \in \mathbb{R}, \quad \lambda \in \lambda \\
& \tfrac{i}{D}\vartheta \leq \text{Tail}_{\frac{i}{D}}(\mathbf{r}^\intercal\boldsymbol{\lambda}) - \widehat{\tau}_i \qquad \forall i = 1, \ldots, D.
\end{aligned}
\tag{3.8}
$$

We refer to this formulation as *RMZ*. As Roman et al. (2013) discuss, the "unscaled" *RMZ* model (3.7) often outputs the portfolio that improves most on the worst outcome of the benchmark distribution (i.e., the left tail). The *RMZ* model (3.8) improves accordingly based on the weights of position, indexed by $i/D,\ i = 1, \ldots, D$. On average, the *RMZ* show some advantages over the un-scaled counterpart from a theoretical and practical point of view. Interested readers are referred to Fábián et al. (2011).

It should also be noted that both risk-based models are never infeasible. They always provide a solution that is SSD efficient, the extreme case being where the benchmark itself is SSD efficient.

To conclude this section, the SSD optimization above is an effective and convenient approach to account for risks in decision-making under uncertainty. However, there is one drawback. The optimal solution depends crucially on the assumption: the population joint CDF of returns is *known* and discrete; moreover, each discrete event has equal probability. In practice, we only observe series of historical data, instead of knowing the population distribution of random variables; moreover, events may not be equally likely to happen, though observations in the sample can be viewed as a discrete realization from the population. These questions are out of OR's scope and then largely ignored in the OR community. But it is relatively straightforward in econometrics research, as we will see in the next section.

3.2.3 Posterior probability of return distribution

The joint CDF $F(\cdot)$ of base assets' return is unknown and has to be estimated using a longitudinal data of historical return, $\mathbf{R}_t = (R_{1t}, R_{2t}, \ldots, R_{nt})^\intercal \sim F(\cdot),\ \forall t = 1, \ldots, T$. To estimate it, we need one assumptions:

Assumption. (Serially IID) for any $t \neq s$, $\text{cov}(R_{it}, R_{js}) = 0,\ i, j = 1, \ldots, n$

The cross-sectional correlation is left unrestricted here. Under the assumption, we present a nonparametric Bayesian method which provides the posterior distribution of $F(\cdot)$. This approach is appealing in two aspects. First, its nonparametric property is compatible with SSD constraints, modeling the true CDF $F(\cdot)$ as an infinite-dimensional parameter. Second, the probabilistic nature of Bayesian methods enables us to consider parameter uncertainties within an optimization scheme, which will be discussed later. A particular nonparametric Bayesian method, called "Bayesian bootstrap" (BB), is well suited for this application. Not only does it efficiently compute the posterior distribution, but its posterior distribution is discrete. The discreteness is particularly helpful here as it is the key assumption SSD optimization used for various formulations. BB method, originally proposed by Rubin (1981), implements the Dirichlet process (DP) model of Ferguson (1973, 1974) by using an improper and noninformative prior.

A general DP model is described as

$$\begin{aligned} F &\sim \text{DP}\big(\alpha(\cdot)\big) \\ F \mid \mathbf{R} &\sim \text{DP}\big(\alpha(\cdot) + \textstyle\sum_{t=1}^{T} \delta_{R_t}(\cdot)\big) \end{aligned}$$

where $\text{DP}(\cdot)$ is a Dirichlet process, a distribution over distributions. The parameter $\alpha(\cdot)$ includes a prior guess at F, say F_0, and reflects how concentrated the prior is around F_0. The function $\delta_{X_i}(\cdot)$ is the measure giving mass one to the point X_i.

Rubin (1981) sets $\alpha \longrightarrow 0$ and calls it the Bayesian bootstrap (BB). It has two advantages. First, the prior tends to be noninformative and does not involve the guess or subjective belief from decision makers. Second, the method becomes very scalable since it is much easier to draw from a finite-dimensional Dirichlet distribution than from a stochastic process like DP. Then, the particular inference we used for the return distribution can be expressed as

$$\begin{aligned} F_T(v) &= \sum\nolimits_{t=1}^{T} \mathbb{1}(X_t \le v)\, p_t \\ (p_1, \ldots, p_t, \ldots, p_T) &\sim \text{Dirichlet}(\mathbf{1}_{T\times 1}) \end{aligned} \tag{3.9}$$

where v is a real value in $\mathbb{R}$, p_t is the posterior probability for $X = X_t$, and Dirichlet$(\cdot)$ is the Dirichlet distribution. Figure 3.1 provides an illustration for the posterior distribution of a standard Gaussian sample of size 100.

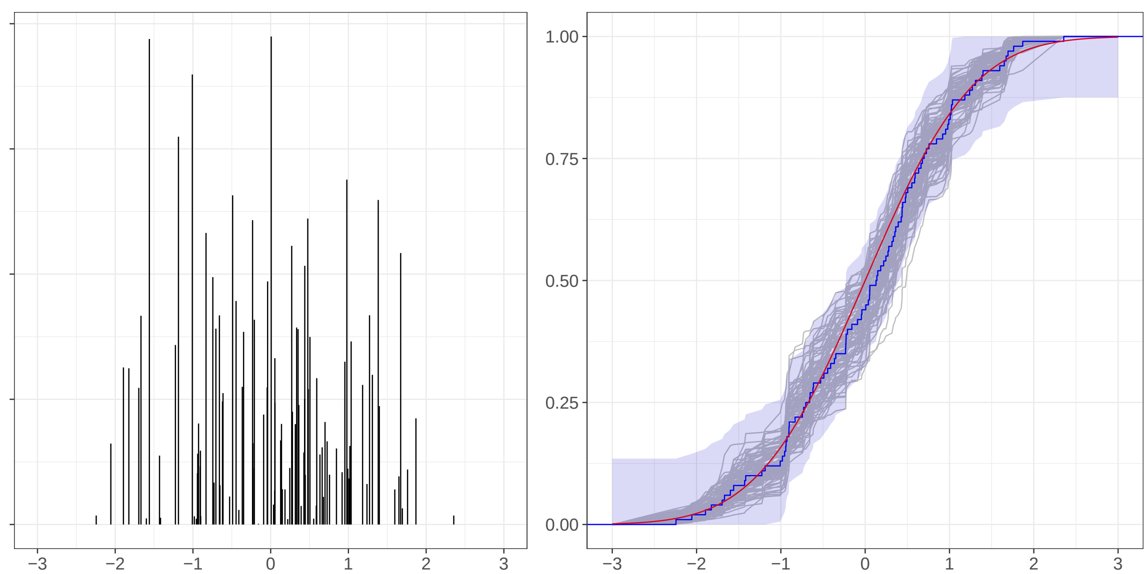


Figure 3.1: *Left*: Discrete probability function from a single draw from Dirichlet$(\cdot)$. *Right*: Population CDF (in red), EDF (in blue), 100 posterior draws (in dark gray), and 95% uniform credible band for the CDF based on 100 posterior draws.

In the portfolio case, $X \in \mathbb{R}^1$ in BB becomes the vector of $\mathbf{R} \in \mathbb{R}^n$. Under the assumption above, we can compute $P(\mathbf{R} = \mathbf{r}_t) = p_t$ in a similar way of computing

$P(X = X_t)$. Now, p_t means the probability for all cross-sectional observations at time t. By doing it, we can keep all the cross-sectional information in our inference.

Therefore, an estimate on $F(\cdot)$ can be expressed by a discrete function $\widehat{F}_T(\mathbf{r}) = \sum_{t=1}^{T} \mathbb{1}(\mathbf{R}_t \leq \mathbf{r})\, p_t$. Here, we relax the restriction of equally-likely event in empirical distribution function and set $P(\mathbf{R} = \mathbf{r}_t) = p_t$, instead of T^{-1}. To estimate two CDFs of interest, that is, $F_Y(\cdot)$ for the benchmark and $F_X(\cdot)$ for new portfolio, we can write

$$\begin{aligned} P(X = \mathbf{r}_t^{\mathsf{T}} \boldsymbol{\lambda}) &= p_t, \qquad t = 1, \ldots, T \\ P(Y = y_k) &= q_k, \qquad k = 1, \ldots, D \end{aligned} \tag{3.10}$$

where y_k is the ordered return rate of Y (i.e., $y_1 \leq \cdots \leq y_k \leq \cdots \leq y_D$), and D is number of support points of Y. An extreme case is that Y has all distinct returns in T periods and then $D = T$; otherwise, we can expect $D \leq T$ in all cases. We can also imply there exists a mapping from p_t to q_k

$$q_k = \sum_{t=1}^{T} \pi_{tk}\, p_t, \quad k = 1, \ldots, D$$

where $\pi_{tk} \equiv P(Y = y_k \mid \mathbf{R} = \mathbf{r}_t)$, an indexing for ranking time-series return Y. Thus, we have $\pi_{tk} \in \{0, 1\}, \forall\, t, k$ and $\sum_{k=1}^{D} \pi_{tk} = 1$, for $t = 1, \ldots, T$. We have shown that the joint CDF $F(\cdot)$ can be characterized by $\mathbf{p} = (p_1, \ldots, p_T)$, and the probability vector $\mathbf{p}$ can further describe distributions of any new portfolio X and the benchmark Y.

3.2.4 Mixed Integer Linear Programming for probabilistic constraints

The SSD optimization formulations in Section 3.2.2 work well when T is "large." It is because the population distribution function is assumed *known* there. Under the serial IID assumption, the empirical distribution function (EDF) is a statistically consistent nonparametric estimator of the CDF. Therefore, it is reasonable for the formulations above to assume the EDF is the population distribution.[6]

However, it could be very inaccurate if the time series is short. Post and Karabati (2016) propose to use an empirical likelihood (EL) method for improving the estimation accuracy. In particular, they replace the equal probabilities $1/T$ in the EDF by the probabilities implied by a set of moment conditions for common risk factors in the EL framework. Here we discuss another data-driven idea that builds upon the Bayesian thinking and optimization with probabilistic (or, chance) constraints.

In a typical chance-constrained optimization problem, decision makers are interested in satisfying a constraint, which involves random variable(s), by a pre-specified probability $1-\alpha$. In our case, we can write

$$\begin{aligned} \max_{\boldsymbol{\lambda}} \quad & f(\boldsymbol{\lambda}) \\ \text{subject to} \quad & \mathbb{P}\big\{\, X(\boldsymbol{\lambda}) \ \mathrm{SD}_2 \ Y \,\big\} \geq 1-\alpha \\ & \boldsymbol{\lambda} \in \Lambda \end{aligned} \tag{3.11}$$

where $\mathbb{P}(\cdot)$ represents the probability measures on the second-order stochastic dominance. Other notations are same as in Equation (3.3). The probabilistic constraints restrict the feasible set of $\boldsymbol{\lambda}$ into the region where the probability of $X(=\mathbf{r}^\intercal\boldsymbol{\lambda})$ second-order dominating Y is not less than $1-\alpha$.

Zhuo (2017) proposes a nonparametric Bayesian method for providing probabilistic measurement on a pair-wise SD of any order. In particular, the posterior probability of the second-order SD relation, like $X(\boldsymbol{\lambda})$ SD_2 Y, is equal to the fraction of posterior drawings where SD relation holds true. We can also apply similar logic to the optimization problem here: in each draw from the posterior for $\widehat{F}(\cdot)$, all inequality constraints are checked. The pre-specified level $1-\alpha$ can be quantified by requiring

[6] Homem-de Mello and Bayraksan (2014) extensively discuss its validity from convergence rates of the optimal value and solutions.

a least $1-\alpha$ proportion of cases when the asset weights $\boldsymbol{\lambda}$ satisfy the constraints.

This idea can be formulated based on Equation (3.4) in a Mixed Integer Linear Program (MILP) as

$$\begin{array}{llcll}
\max\limits_{\boldsymbol{\lambda},\, s_{kt},\, z^b} & \mathbb{E}[\mathbf{r}^{\intercal}\lambda] & & & \\
\text{s.t.} & \sum\limits_{i=1}^{n}\lambda_i r_{it}+s_{kt} & \geq & y_k & k=1,\dots,D,\, t=1,\dots,T\\
& s_{kt} & \geq & 0 & k=1,\dots,D,\, t=1,\dots,T\\
& \sum\limits_{t=1}^{T}p_t^b s_{kt}-z^b M & \leq & F_2^b(Y;y_k), & k=1,\dots,D,\, b=1,\dots,B\\
& z^b & \in & \{0,1\}, & b=1,\dots,B\\
& \lambda_i & \geq & 0 & i=1,2,\dots,n\\
& \sum\limits_{b=1}^{B} z^b & \leq & \alpha B & \\
& \sum\limits_{i=1}^{n}\lambda_i & = & 1 &
\end{array} \tag{3.12}$$

where $F_2^b(Y;y_k)=\mathbb{E}_b\big[(y_k-Y)_+\big]=\sum_{j=1}^{D}q_j^b(y_k-y_j)_+$, z^b is an indicator function which keeps track if the SSD constraints is met. M is an infinitely large constant, such that the optimal solution for z^b takes value 0 if $X\,\mathrm{SD}_2\,Y$; otherwise takes value 1. By considering this formulation, the set of our decision variables expands to $\{\boldsymbol{\lambda}, s_{kt}, z_b\}$ here from $\{\boldsymbol{\lambda}, s_{kt}\}$ in Equation (3.4); the number of constraints also increases considerably. Though it looks like a computational burden, commercial solvers like Ilog CPLEX can handle this size optimization skillfully.

3.3 Empirical study

This section examines and compares two major SSD optimization formulations (i.e., *DR* and *RMZ*), using two data sets from the U.S. stock markets. We aim to construct a portfolio over n available assets according to the SSD criteria. That is, we must decide how much of each asset should be invested in the portfolio constructed by specific optimization formulations. The purpose is to compare the performance of different SSD-efficient portfolio constructions based on the same data.

3.3.1 Data and investment strategy

Two data sets we used are the daily industry portfolios return from *Fama & French Data library*, and the weekly stock returns from the Center of Research in Security Prices (CRSP). The data can be used as input for several portfolio construction methods.

Our empirical implementation is based on a rolling-window scheme, as shown in Figure 3.2. That is, the investment strategy is to re-balance the portfolio after each block of a fixed number of time periods. For example, we compute portfolio weights using a rolling in-sample window of 250 return observations. We initially set the in-sample window on the first 250 trading days and solve the model for the optimal portfolio weights, which would be used to select assets for next trading days. Then we evaluate the performance of the selected portfolio on the following 60 (out-of-sample) trading days. Next, we update the in-sample window by including the 60 previously out-of-sample periods and removing the first 60 periods from the in-sample window. We then re-balance the portfolio by re-solving the model, and repeat until the end of the data set.

3.3.2 Performance measures

The out-of-sample performance of a portfolio construction is often evaluated by a number of performance measures. In this study, we choose the following six performance metrics widely used in the finance literature (e.g., see DeMiguel et al. 2007).

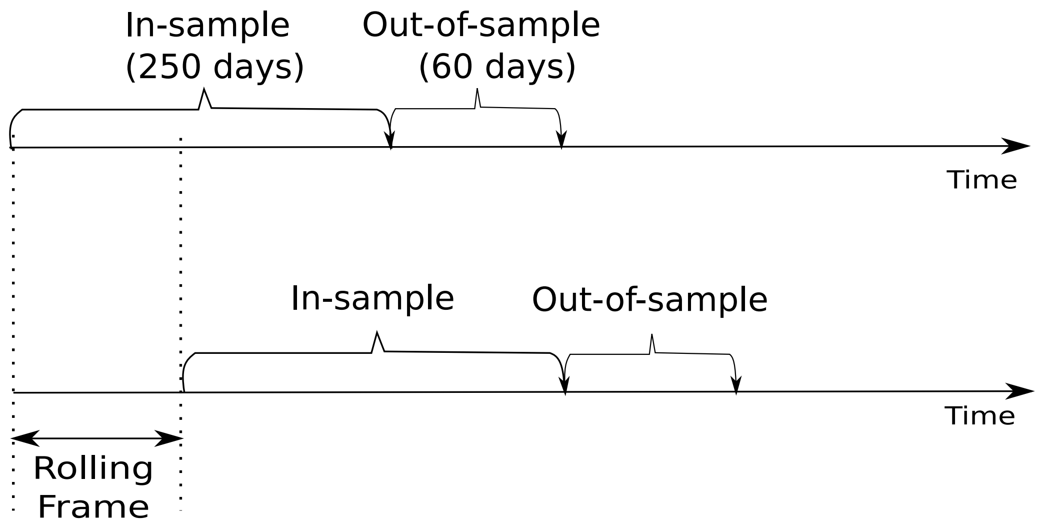


Figure 3.2: Scheme of the rolling time window in the Fama–French industry dataset

For notation, we denote the out-of-sample portfolio return by R^{out} and set a constant risk-free rate of return $r_f = 0$.

- *Max Drawdown* (MDD) is the maximum loss from a peak to a trough of a portfolio, before a new peak is attained. It is an indicator of downside risk over a specified time period:

$$\text{MDD}(T) = \max_{0 \leq \tau \leq T} \left(\max_{0 \leq t \leq \tau} V(t) - V(\tau) \right)$$

  where $V(\cdot)$ is the value of portfolio. The smaller the value is, the better the portfolio performance is.

- *Sharpe Ratio* (Sharpe, 1966, 1994) is the ratio between the average of $R^{\text{out}-r_f}$ and its standard deviation, that is,

$$\text{Sharpe Ratio} = \frac{\mathbb{E}(R^{\text{out}} - r_f)}{\sigma(R^{\text{out}})}.$$

It is one of the most important reward-vesus-risk ratios. It reflects how much excess return is earned for every unit of risk exposure. The larger is the ratio, the better is the portfolio performance.

- *Sortino Ratio* (Sortino and Price, 1994) is the ratio between the average of of $R^{\text{out}} - r_f$ and the downside deviation, that is,

$$\text{Sortino Ratio} = \frac{\mathbb{E}(R^{\text{out}} - r_f)}{\sigma\big(\min(R^{\text{out}} - r_f, 0)\big)}.$$

  It singles out the "bad" risk exposure from all risks. The larger the value is, the better the portfolio performance is.

- *Ulcer Index* (UI) is another measure of volatility in the downward direction:

$$R^{\%} = 100 \times \frac{R^{\text{out}} - \max(R^{\text{out}})}{\max(R^{\text{out}})},$$
$$UI = \sqrt{N^{-1}\Big(R_1^{\%} + \cdots + R_T^{\%}\Big)}.$$

  The larger the value is, the better the portfolio performance is.

- *Omega Ratio* (Keating and Shadwick, 2002) is a risk-return performance measure based on the distribution function of R^{out}, say F^{out}:

$$\Omega(r) = \frac{\displaystyle\int_r^{+\infty} \big(1 - F^{\text{out}}(v)\big)\, dv}{\displaystyle\int_{-\infty}^{r} F^{\text{out}}(v)\, dv}$$

  where we set $r = 0$ in our empirical study for the sake of convenience. The larger the value is, the better the portfolio performance is.

- *Win Ratio* is the fraction of time periods with positive excess return:

$$\text{Win Ratio} = \frac{\sum_{t=1}^{T} \mathbb{1}(R_t^{\text{out}} \geq r_f)}{T}.$$

The larger the value is, the better the portfolio performance is.

For each data set, we will report these six performance metrics along with the common statistics (i.e., the average daily return, the compound annual growth rate, standard deviation, skewness and kurtosis), where the best results are marked in bold.

3.3.3 Fama & French industry portfolios

In this data set, the base assets include a set of 49 industry portfolios, which are formed by grouping individual stocks listed on NYSE, AMEX and NASDAQ markets by their four-digit Standard Industry Classification (SIC) codes. We study the daily return of these base assets from 01/03/2000 to 12/30/2016, with 250 trading days for estimation and 30 days for holding, and re-balance the portfolio at the end of each holding period.

The benchmark here is the equal-weighted (EW) average of the base assets (i.e., the naive 1/N portfolio). Such heuristic diversification is a simple but effective way to achieve robust performance. DeMiguel et al. (2007) claim the EW is a good choice for the benchmark given the fact that EW often outperforms many "optimal" methods like the MV method in terms of out-of-sample return. Since the base assets are already diversified industry portfolios, we can expect there should be no consistent concentrated position in a single "asset".

Table 3.2 summarizes the out-of-sample performance of the competing portfolios. The benchmark ('EW') on average yields 17.25% per annum with a standard deviation of 1.20 percentage points in the sample period. The negative skewness shows the higher chance of losses than of gains. That is, the downside risk is not reduced by heuristic diversification. Besides, it is clearly dominated by the SSD-enhanced portfolio strategy for most of the performance metrics, except the win ratio.

The performance enhancement from *DR* and *RMZ* is significant. The *DR* im-

Table 3.2: Performance in Fama & French industry data set

Metrics	DR	RMZ	EW
Avg. Daily Ret. Rate	0.0746%	**0.0810%**	0.0686%
Comp. Ann. Growth Rate	19.19%	**21.46%**	17.25%
Std. Dev.	0.0120	0.0112	0.0120
Skewness	-0.1609	-0.3028	-0.3164
Kurtosis	4.9046	5.4569	7.2037
Max Drawdown	45.98%	48.60%	59.01%
Sharpe Ratio	0.9880	**1.1428**	0.9096
Sortino Ratio	1.6135	**1.8422**	1.4397
Ulcer Index	9.9332	9.3326	12.0146
Omega Ratio	1.1928	**1.2255**	1.1816
Win Ratio	54.93%	55.87%	**56.28%**

Note: **value in bold** is the best performer (highest in row).

proves the average annual return by 1.67 percentage points to 19.19%, and the *RMZ* by 4.21 points to 21.46%, respectively. *RMZ* reduces the standard deviation by 0.08 percentage points. Though *DR* has same standard deviation as the benchmark EW, it is less left-skewed as its skewness is -0.1609, less than -0.3164 of EW. Both *DR* and *RMZ* score better for all performance metrics, except the win ratio. EW has the highest win ratio at 56.28%, which reflects the return of EW portfolio may be very volatile. Other than this, our SSD-enhanced portfolios have a better ratio of reward over risk.

Figure 3.3 shows the development of the dollar value of three portfolios over the entire sample period. The initial investment in the first period in every portfolio is one dollar. The portfolios are formed and rebalanced at the beginning of a 30-trading-day holding period based on a trailing 250-trading-day estimation window of daily returns. The first estimation window is 01/03/2000 - 12/28/2000 and first holding period is 12/28/2000 - 02/12/2001. The top panel illustrates the cumulative performance of the competing portfolios for the entire sample period. Not surprisingly, all three portfolios have a similar trend and pattern. Among them, *RMZ* ranks above all other portfolio most of time, while *DR* is very close to *RMZ* until year 2011. The second and third

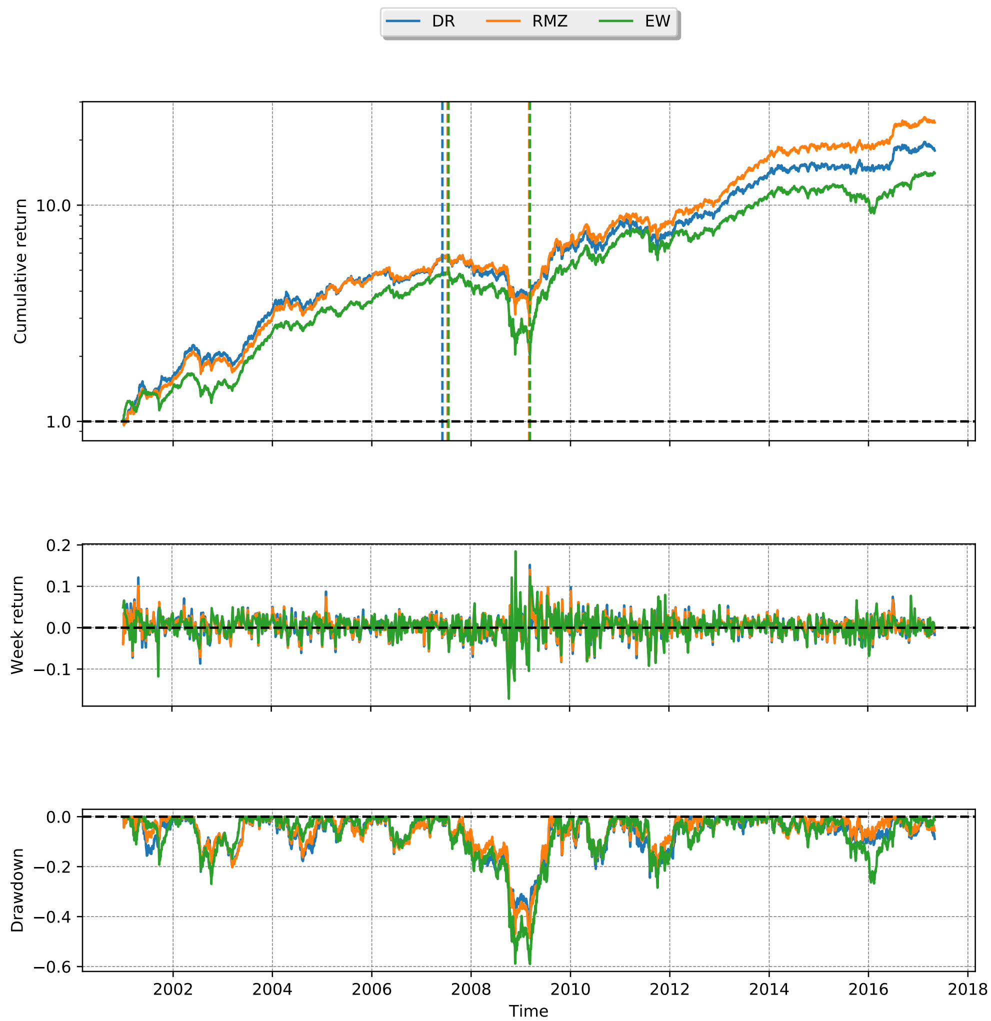


Figure 3.3: Out-of-sample performance in the Fama & French data set

panels show the benchmark is very volatile and has larger drawdown for most of the time. To conclude, both DR and RMZ show good signs of improvement over EW.

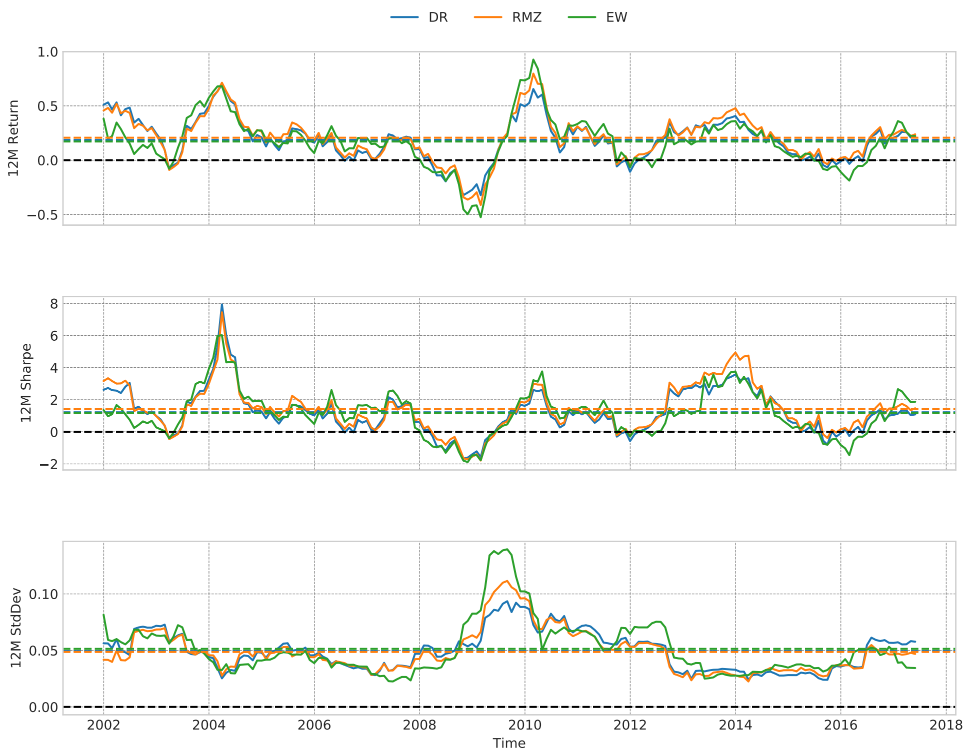


Figure 3.4: 12 months rolling returns in the Fama & French data set

Figure 3.4 presents the rolling return, Sharpe ratio and standard deviation of three portfolios. The rolling returns is a more realistic way of looking at investment returns, which provide a dynamic look at each data point of re-balance. The first rolling return is the annualized average return for a period spanning from 12/28/2000 to 12/28/2001. The second rolling return is computed after 30 trading days. The rolling returns among the three portfolios are similar, but DR and RMZ have a higher Sharpe ratio since their standard deviations are relatively small.

3.3.4 Standard & Poor's 500

The data used here contains weekly returns of 442 base assets over 595 weeks, from 11/2004 to 04/2016. The base assets are individual stocks with more than ten years of observations. A natural benchmark is the Standard & Poor's 500 (SP500). *RMZ* and *DR* formulations are applied to construct the portfolio that enhances SP500. Different from previous empirical study, we use 52 weeks for estimation and 12 weeks for holding, and keep rebalancing every 12 weeks.

Table 3.3 summarizes the out-of-sample performance of the competing portfolios. The benchmark ('EW') on average yields 4.92% per annum with a standard deviation of 1.21 percentage points in the sample period. The negative skewness shows the higher chance of losses when compared to of gains. That is, the downside risk is not reduced by heuristic diversification. Besides, it is clearly dominant by SSD-enhanced portfolio strategy for all performance metrics.

Table 3.3: Performance in SP500 data set

Metrics	**DR**	**RMZ**	**SP500**
Avg. Daily Ret. Rate	**0.0603%**	0.0593%	0.0264%
Comp. Ann. Groth Rate	10.00%	**13.34%**	4.92%
Std. Dev.	0.0205	0.0149	0.0121
Skewness	-0.0818	0.2395	-0.0539
Kurtosis	39.1423	39.4033	36.3180
Max Drawdown	66.32%	**49.57%**	56.43%
Sharpe Ratio	0.4660	**0.6308**	0.3462
Sortino Ratio	0.6894	**0.9445**	0.5062
Ulcer Index	28.8132	**15.7746**	18.2660
Omega Ratio	1.2042	**1.2916**	1.1520
Win Ratio	54.42%	**56.83%**	54.61%

Note: **value in bold** is the best performer (highest within row).

RMZ is a clear winner, though both SSD formulations provide significant performance enhancement. The *DR* improves the average return by 5.08 percentage points to 10.00%, and *RMZ* by 8.42 points to 13.34%. Their higher standard deviations may be due to their portfolio having more upward risk exposure. It is especially true for

RMZ whose skewness is positive. Such preferable asymmetric return profile for RMZ results in its advantages over DR in most of the performance metrics. For example, the factor that RMZ has highest Sortino ratio can reflect its superior management on the downside risk. Interestingly, the DR is supposed to control the downside risk, but it has a larger Max Drawdown and Ulcer index than EW, though its Sharpe and Sortino ratios are still higher.

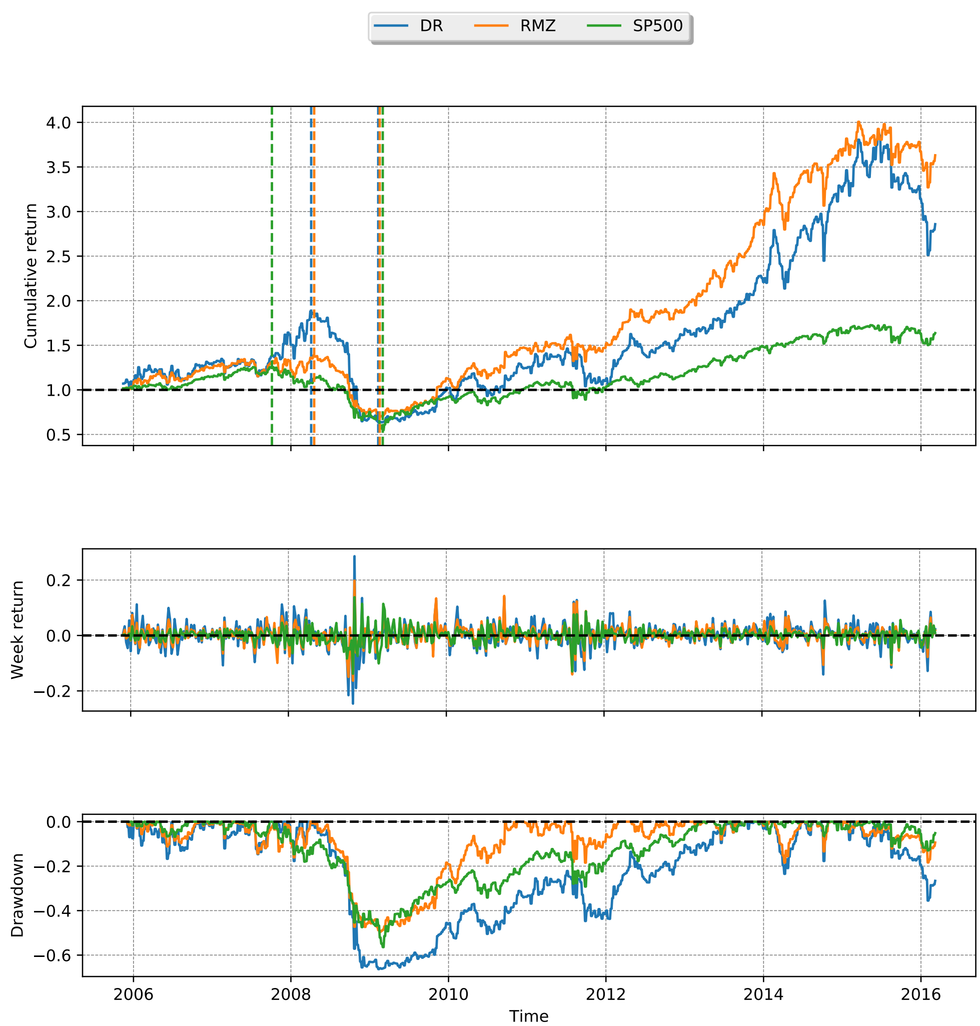


Figure 3.5: Out-of-sample performance in the SP500 data set

Figure 3.5 shows the dynamic development of the dollar value of three portfolios over the entire sample period. At the end of period, top panel shows RMZ and DR outperform the SP500 by a large margin. The bottom panel further reveals that DR has the largest peak-to-trough decline in the value of its portfolio, which suggests its largest standard deviation is mainly associated with the bad risk.

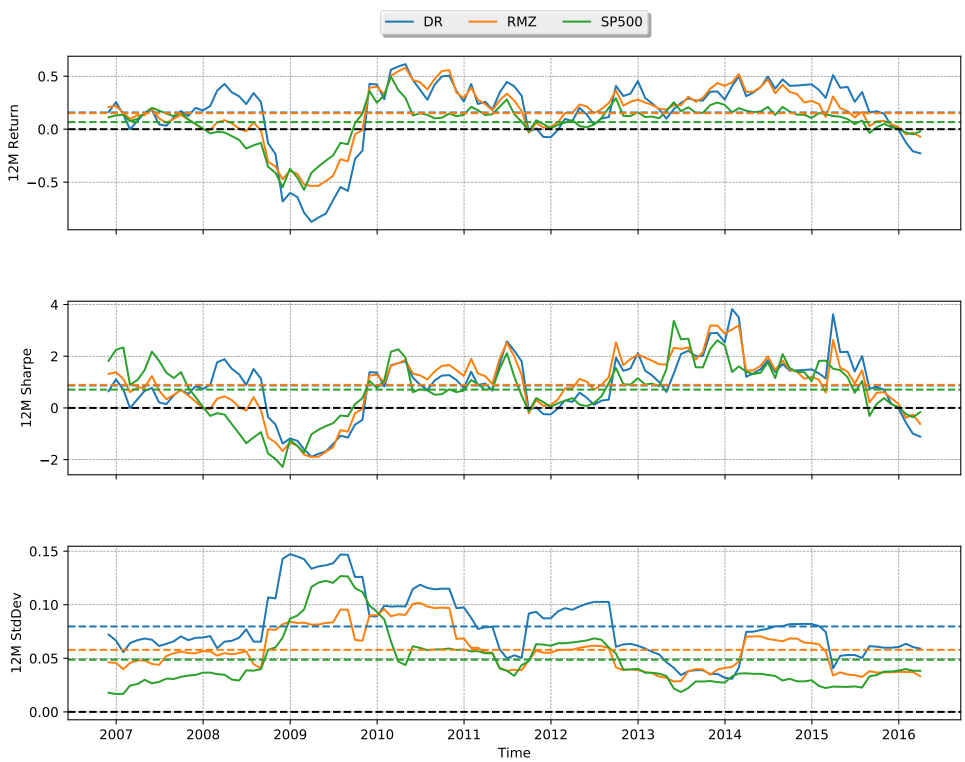


Figure 3.6: 12 months rolling returns in the SP500 data set

Figure 3.6 presents the rolling return, Sharpe ratio and standard deviation of three portfolios. DR has on average similar rolling return like RMZ, but its average standard deviation is much higher. Therefore, the rolling Sharpe ratio for DR is lower than RMZ, but still higher than SP500. To conclude, the SSD-efficient portfolios are more profitable and/or safer than the popular index SP500.

3.4 Conclusion

Stochastic dominance is an effective way to enhance a benchmark portfolio. It is also practical thanks to rapidly developing computing technology. However, some issues need to be addressed for their better performance. First, we need to make sure the optimization formulation can successfully select a portfolio that stochastically dominates the benchmark at the second order. Second, the optimization may be sensitive to estimation errors. We propose an idea to solve this question, which combines the probabilistic nature in the Bayesian inference and mixed-integer linear programming to construct a feasible set of dominant portfolios at some confidence level. Further research can focus on exploring an effective way to implement the optimization formulation we propose here.

We contribute to the portfolio optimization literature by developing a framework to incorporate statistical uncertainty in the sample. We also hope to contribute to the stochastic optimization literature by showing that Bayesian inference may provide insight into the study of data-driven chance constraints.

Appendix A

Technical Appendix to Chapter 3

A.1 More formulations for SSD optimization

Luedtke (2008) develops two new formulations for optimization under SSD constraints. He claims these two formulation gain a huge reduction in terms of size of the problem, and therefore are faster and efficient. Two approaches are based on a different idea, which is to find another variable W such that $X_t = \mathbf{r_t}^\intercal \boldsymbol{\lambda} \geq W_t,\ \forall t$ and W SD_2 Y. Assume $\pi_{tk} = \mathrm{prob}(Y = y_k \mid W = w_t)$

$$\begin{array}{lrcll}
\max\limits_{\boldsymbol{\lambda}} & \mathbb{E}(r)^\intercal \boldsymbol{\lambda} & & & \\
\text{s.t.} & \displaystyle\sum_{k=1}^{D} \pi_{tk} & = & 1 & t = 1, \ldots, T \\
& \displaystyle\sum_{t=1}^{T} p_t \pi_{tk} & = & q_k, & k = 1, \ldots, D \\
& \displaystyle\sum_{i=1}^{n} r_{it} \lambda_i & \geq & \displaystyle\sum_{k=1}^{D} y_k \pi_{tk} & t = 1, \ldots, T
\end{array} \tag{A.1}$$

We refer it to the formulation above *L1*, the formulation below *L2*.

$$
\begin{aligned}
\max_{x} \quad & \mathbb{E}(r)^{\intercal}\boldsymbol{\lambda} & & & \\
\text{s.t.} \quad & \sum_{k=1}^{D}\pi_{tk} &=& \ 1 & t=1,\ldots,T \\
& \sum_{i=1}^{n} r_{it}\lambda_i &\geq& \sum_{k=1}^{D} y_k\pi_{tk} & t=1,\ldots,T \\
& v_k - \sum_{t=1}^{T} p_t\pi_{tk} &=& \ 0 & k=1,\ldots,D \\
& \sum_{j=1}^{k-1} v_j(y_k-y_j) &\leq& \sum_{j=1}^{k-1} q_j(y_k-y_j) & t=1,\ldots,T
\end{aligned}
\tag{A.2}
$$

These two linear programming for SSD optimization have only $O(N+D)$ constraints, as opposed to $O(ND)$ constraints in the formulation *DR*. It is supposed to yield a huge improvement in solution time, especially for cases in which $N=D$.

A.2 Simulation

This purposes of this simulation study is two-fold; first, it illustrates and compares all SSD optimization formulation in one same setting, with emphasize on sanity check of within sample SD_2, and second, it computes and compares the in-sample cumulative returns / values by adopting the asset weights each formulation suggests. For the sake of convenience, we list the competing models in table A.1.

Here is a brief description of DGP. We draw 100 observations on 10 assets from a joint distribution, which is characterized by a multivariate t distribution with zero means and non-diagonal covariance matrix. The benchmark portfolio Y is constructed by investing 10 assets equally. We implement both the MV and SSD-efficient approach to compute the optimal asset weights respectively.

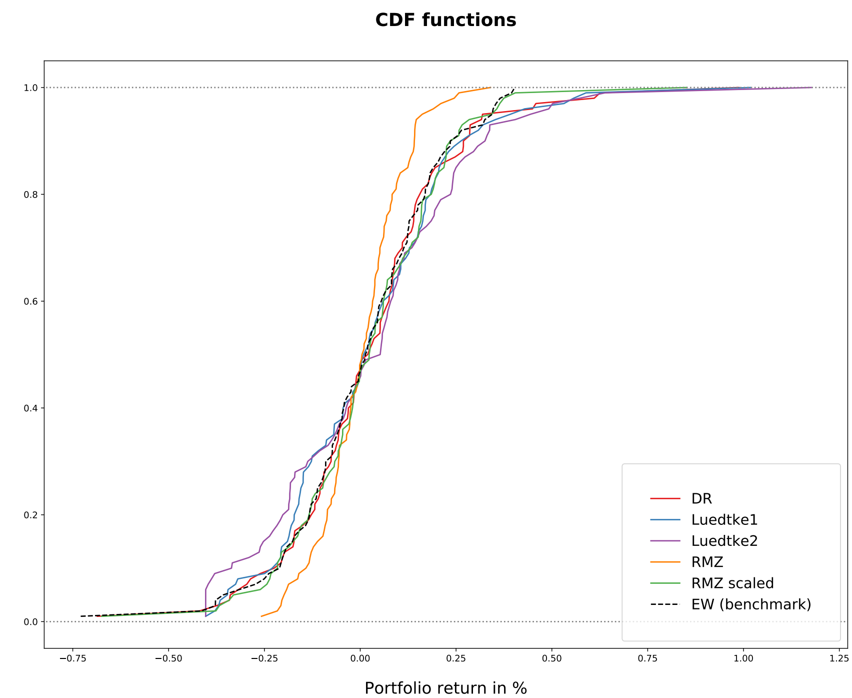
In Figure A.1a, we plot the distribution functions for benchmark (dash line) as

Table A.1: Portfolio selection models

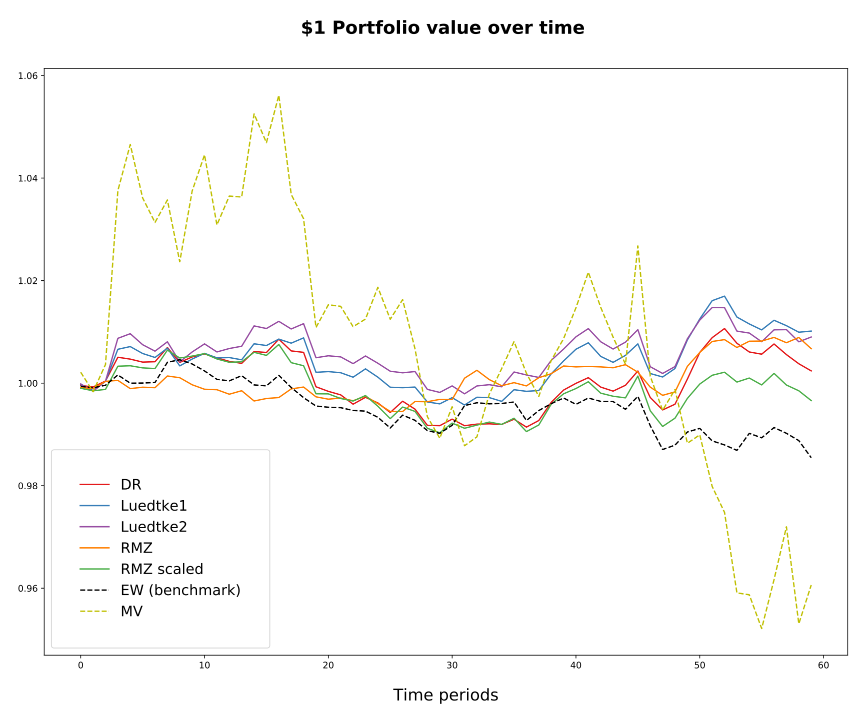
Model Name	**Formulation**	**Reference**
DR	(3.4)	Dentcheva and Ruszczyński (2006)
L1	(A.1)	cSSD1 in Luedtke (2008)
L2	(A.2)	cSSD2 in Luedtke (2008)
RMZ	(3.7)	Roman et al. (2013)
RMZ scaled	(3.8)	Roman et al. (2013)
MV	-	Markovic's Mean-Variance method
EW	-	Equal weights method

well as other portfolios constructed. A visual check tells that *DR, RMZ* and *RMZ scaled* work, while L1 and *L2* do not. In particular, *RMZ* is very conservative in the sense that it has the smallest variance. That is to say, it sacrifices rewards for minimal risks. *RMZ scaled* is slightly better job for a reward-risk tradeoff. *DR* performs best in this simulation; it controls the bad risk as good as *RMZ scaled*, while it has a larger good risk exposure. On the other side, the portfolios constructed by L1 and *L2* are not dominant over the benchmark w.r.t. SSD. They are more volatile than the benchmark, though they strongly control the outcome of extremely bad scenario.

Figure A.1b studies, if one dollar is invested, what the ultimate the return we can have by each strategy. The MV approach is the worst, not only has the lowest value at the end, but very volatile. All SSD-based strategy perform better than the benchmark. Though *L1* and *L2* rank the top two, we still exclude them in the main body and attribute their success to lucks.



(a) Sanity check on the validity of SSD optimization formations



(b) Return comparisons among the benchmark, MV and SSD-efficiency

Figure A.1: Simulations

A.3 Dynamics of asset weights in SP500

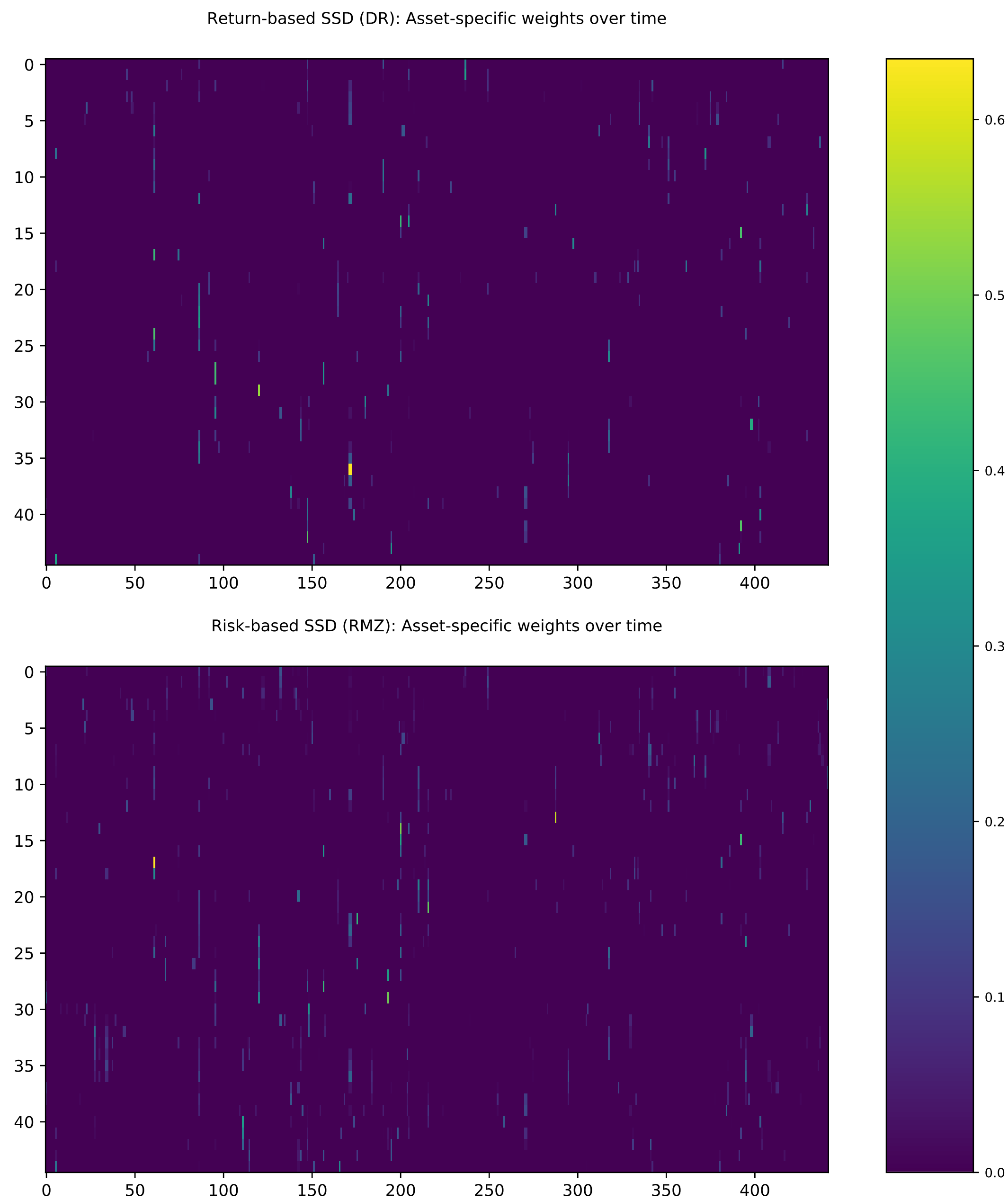


Figure A.2: How often do asset weights update?

Bibliography

Abul Naga, R. H. and T. Yalcin (2008). Inequality measurement for ordered response health data. *Journal of Health Economics 27*(6), 1614–1625.

Aldor-Noiman, S., L. D. Brown, A. Buja, W. Rolke, and R. A. Stine (2013). The power to see: A new graphical test of normality. *The American Statistician 67*(4), 249–260.

Allison, R. A. and J. E. Foster (2004). Measuring health inequality using qualitative data. *Journal of Health Economics 23*(3), 505–524.

Andrews, D. W. K. and P. J. Barwick (2012). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica 80*(6), 2805–2826.

Andrews, D. W. K. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*(1), 119–157.

Avramov, D. and G. Zhou (2010). Bayesian portfolio analysis. *Annu. Rev. Financ. Econ. 2*(1), 25–47.

Bago d'Uva, T., E. Van Doorslaer, M. Lindeboom, and O. O'Donnell (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics 17*(3), 351–375.

Banks, D. L. (1988). Histospline smoothing the Bayesian bootstrap. *Biometrika 75*(4), 673–684.

Baron, D. P. (1977). On the utility theoretic foundations of mean-variance analysis. *The Journal of Finance 32*(5), 1683–1697.

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica 71*(1), 71–104.

Bawa, V. S. (1975). Optimal rules for ordering uncertain prospects. *Journal of*

Financial Economics 2(1), 95–121.

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

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*(3), 414–427.

Canay, I. A. and V. Kamat (2017). Approximate permutation tests and induced order statistics in the regression discontinuity design. Technical Report CWP21/17, Centre for Microdata Methods and Practice (CeMMAP).

Casella, G. and R. L. Berger (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association 82*(397), 106–111.

Chamberlain, G. and G. W. Imbens (2003). Nonparametric applications of Bayesian inference. *Journal of Business & Economic Statistics 21*(1), 12–18.

Contoyannis, P., A. M. Jones, and N. Rice (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics 19*(4), 473–503.

Crossley, T. F. and S. Kennedy (2002). The reliability of self-assessed health status. *Journal of Health Economics 21*(4), 643–658.

Currie, J. and B. C. Madrian (1999). Health, health insurance and the labor market. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3C, Chapter 50, pp. 3309–3416. Elsevier.

Cutler, D. M. and E. Richardson (1997). Measuring the health of the U.S. population. *Brookings Papers on Economic Activity: Microeconomics 1997*, 217–272.

Davidson, R. and J.-Y. Duclos (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica 68*(6), 1435–1464.

Davidson, R. and J.-Y. Duclos (2013). Testing for restricted stochastic dominance. *Econometric Reviews 32*(1), 84–125.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. Baltimore: The Johns Hopkins University Press.

Deaton, A. and C. Paxson (1998). Health, income, and inequality over the life cycle. In *Frontiers in the Economics of Aging*, pp. 431–462. University of Chicago Press.

DeGroot, M. H. (2004). *optimal statistical decisions.* Wiley.

DeMiguel, V., L. Garlappi, and R. Uppal (2007). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies 22*(5), 1915–1953.

Dentcheva, D. and A. Ruszczyński (2003). Optimization with stochastic dominance constraints. *SIAM Journal on Optimization 14*(2), 548–566.

Dentcheva, D. and A. Ruszczyński (2006). Portfolio optimization with stochastic dominance constraints. *Journal of Banking & Finance 30*(2), 433–451.

Dykstra, R. L. and P. Laud (1981). A Bayesian nonparametric approach to reliability. *Annals of Statistics 9*(2), 356–367.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics 7*(1), 1–26.

Fábián, C. I., G. Mitra, and D. Roman (2011). Processing second-order stochastic dominance models using cutting-plane representations. *Mathematical Programming 130*(1), 33–57.

Fábián, C. I., G. Mitra, D. Roman, and V. Zverovich (2011). An enhanced model for portfolio choice with SSD criteria: a constructive approach. *Quantitative Finance 11*(10), 1525–1534.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics 1*(2), 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics 2*(4), 615–629.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis* (3 ed.). Chapman & Hall/CRC.

Goldman, M. and D. M. Kaplan (2017). Comparing distributions by multiple testing across quantiles or CDF values. Working paper, available at `https://faculty.missouri.edu/~kaplandm`.

Goutis, C., G. Casella, and M. T. Wells (1996). Assessing evidence in multiple hypotheses. *Journal of the American Statistical Association 91*(435), 1268–1277.

Groot, W. (2000). Adaptation and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics 19*(3), 403–420.

Hadar, J. and W. R. Russell (1969). Rules for ordering uncertain prospects. *American*

Economic Review 59(1), 25–34.

Hanoch, G. and H. Levy (1969). The efficiency analysis of choices involving risk. *The Review of Economic Studies 36*(3), 335–346.

Heiss, F. (2011). Dynamics of self-rated health and selective mortality. *Empirical Economics 40*(1), 119–140.

Heiss, F., S. F. Venti, and D. A. Wise (2014). The persistence and heterogeneity of health among older Americans. NBER Working Paper 20306, National Bureau of Economic Research.

Hernández-Quevedo, C., A. M. Jones, and N. Rice (2005). Reporting bias and heterogeneity in self-assessed health. evidence from the British Household Panel Survey. HEDG Working Paper 05/04, Health, Econometrics and Data Group, The University of York.

Homem-de Mello, T. and G. Bayraksan (2014). Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science 19*(1), 56–85.

Jones, A. M., N. Rice, S. Robone, and P. R. Dias (2011). Inequality and polarisation in health systems' responsiveness: a cross-country analysis. *Journal of Health Economics 30*(4), 616–625.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics 16*(2), 163–178.

Kaplan, D. M. and L. Zhuo (2017). Bayesian and frequentist nonlinear inequality tests. Working paper, available at `https://faculty.missouri.edu/~kaplandm`.

Kaur, A., B. L. S. Prakasa Rao, and H. Singh (1994). Testing for second-order stochastic dominance of two distributions. *Econometric Theory 10*(5), 849–866.

Keating, C. and W. F. Shadwick (2002). A universal performance measure. *Journal of Performance Measurement 6*(3), 59–84.

Kline, B. (2011). The Bayesian and frequentist approaches to testing a one-sided hypothesis about a multivariate mean. *Journal of Statistical Planning and Inference 141*(9), 3131–3141.

Kobus, M. and P. Miłoś (2012). Inequality decomposition by population subgroups for ordinal data. *Journal of Health Economics 31*(1), 15–21.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata*

Journal 10(2), 165–199.

Kuosmanen, T. (2004). Efficient diversification according to stochastic dominance criteria. *Management Science 50*(10), 1390–1406.

Lancaster, T. (2003). A note on bootstraps and robustness. Available at SSRN: `https://ssrn.com/abstract=896764`.

Lander, D., D. Gunawan, W. Griffiths, and D. Chotikapanich (2016, May). Bayesian assessment of Lorenz and stochastic dominance using a mixture of gamma densities. Working paper, available at `https://ideas.repec.org/p/mlb/wpaper/2023.html`.

Levy, H. (2015). *Stochastic dominance: Investment decision making under uncertainty*. Springer.

Lindeboom, M. and E. van Doorslaer (2004). Cut-point shift and index shift in self-reported health. *Journal of Health Economics 23*(6), 1083–1099.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005). Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies 72*(3), 735–765.

Linton, O., T. Post, and Y.-J. Whang (2014). Testing for the stochastic dominance efficiency of a given portfolio. *The Econometrics Journal 17*(2), S59–S74.

Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Annals of Statistics 15*(1), 360–375.

Lo, A. Y. (1993). A Bayesian method for weighted sampling. *Annals of Statistics 21*(4), 2138–2148.

Lubrano, M. and A. A. J. Ndoye (2016). Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach. *Computational Statistics & Data Analysis 100*, 830–846.

Luedtke, J. (2008). New formulations for optimization under stochastic dominance constraints. *SIAM Journal on Optimization 19*(3), 1433–1450.

Pástor, L. (2000). Portfolio selection and asset pricing models. *The Journal of Finance 55*(1), 179–223.

Pástor, L. and R. F. Stambaugh (2000). Comparing asset pricing models: an investment perspective. *Journal of Financial Economics 56*(3), 335–381.

Porter, R. B. and J. E. Gaumnitz (1972). Stochastic dominance vs. mean-variance portfolio analysis: an empirical evaluation. *The American Economic Review 62*(3), 438–446.

Post, T. (2003). Empirical tests for stochastic dominance efficiency. *The Journal of Finance 58*(5), 1905–1931.

Post, T. and S. Karabati (2016). Portfolio optimization based on stochastic dominance and empirical likelihood. Unpublished.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhyā: The Indian Journal of Statistics, Series A 27*(2/4), 311–324.

Rao, J. N. K. and C. F. J. Wu (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association 83*(401), 231–241.

Ravallion, M. (1994). *Poverty comparisons, Harwood fundamentals of pure and applied economics*. Harwood Academic Publisher.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (2nd ed.). Springer.

Rockafellar, R. T. and S. Uryasev (2000). Optimization of conditional value-at-risk. *Journal of Risk 2*, 21–42.

Rockafellar, R. T. and S. Uryasev (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance 26*(7), 1443–1471.

Roman, D., K. Darby-Dowman, and G. Mitra (2006). Portfolio construction based on stochastic dominance and target return distributions. *Mathematical Programming 108*(2), 541–569.

Roman, D., G. Mitra, and V. Zverovich (2013). Enhanced indexation based on second-order stochastic dominance. *European Journal of Operational Research 228*(1), 273–281.

Rothschild, M. and J. E. Stiglitz (1971). Increasing risk II: Its economic consequences. *Journal of Economic Theory 3*(1), 66–84.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics 9*(1), 130–134.

Scaillet, O. and N. Topaloglou (2010). Testing for stochastic dominance efficiency. *Journal of Business & Economic Statistics 28*(1), 169–180.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business 39*(1), 119–138.

Sharpe, W. F. (1994). The Sharpe ratio. *Journal of Portfolio Management 21*(1), 49–58.

Solon, G., S. J. Haider, and J. M. Wooldridge (2015). What are we weighting for? *Journal of Human Resources 50*(2), 301–316.

Sortino, F. A. and L. N. Price (1994). Performance measurement in a downside risk framework. *the Journal of Investing 3*(3), 59–64.

van Doorslaer, E. and A. M. Jones (2003). Inequalities in self-reported health: validation of a new approach to measurement. *Journal of Health Economics 22*(1), 61–87.

van Kippersluis, H., T. Van Ourti, O. O'Donnell, and E. van Doorslaer (2009). Health and income across the life cycle and generations in Europe. *Journal of Health Economics 28*(4), 818–830.

Wagstaff, A. and E. Van Doorslaer (1994). Measuring inequalities in health in the presence of multiple-category morbidity indicators. *Health Economics 3*(4), 281–291.

Weng, C.-S. (1989). On a second-order asymptotic property of the Bayesian bootstrap mean. *Annals of Statistics 17*(2), 705–710.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.

Zellner, A. and V. K. Chetty (1965). Prediction and decision problems in regression models from the Bayesian point of view. *Journal of the American Statistical Association 60*(310), 608–616.

Zhuo, L. (2017). Nonparametric Bayesian inference on stochastic dominance. Working paper.

VITA

Longhao Zhuo was born and raised in Kunming, China in 1988. He received his Bachelor Degree in Statistics in 2011 and Master Degree in Finance Engineering in 2013, both from the Huazhong University of Science and Technology. In August of 2013, he enrolled as a graduate student at the University of Missouri.

He continued to pursue his Doctorate Degree in Econometrics and Quantitative Economics under the advise of Professor David M. Kaplan.