Machine Learning & Portfolio Optimization

Gah-Yi Ban

NUS-USPC Workshop on Machine Learning and FinTech
Nov 2017

Portfolio Optimization

Consider the portfolio optimization problem (Markowitz, 1952):

$$\begin{array}{lll} \min\limits_{\mathbf{w}\in\mathbb{R}^p} & \mathbf{w}^\top \Sigma \mathbf{w} & \\ s.t. & \mathbf{w}^\top \mu = R & \\ & \mathbf{w}^\top \mathbf{1} = 1 & \end{array} \tag{MV}$$

where

- $\mathbf{X}$: $p \times 1$ random vector of relative returns
- $\mu = E(\mathbf{X})$: mean returns
- $\Sigma = Cov(\mathbf{X})$: $p \times p$ covariance matrix for the relative returns
- Solution: $\mathbf{w}_0(R)$
- Same if return constraint is relaxed to $\mathbf{w}^\top \mu \geq R$

Sample Average Approximation

- In practice, we don't know the distribution P of $\mathbf{X}$ but have data
- Suppose we have n iid observations of asset returns from P: $\mathcal{X}_n = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$.
- Then solve

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^p} \quad & \mathbf{w}^\top \hat{\Sigma}_{1:n} \mathbf{w} \\ s.t. \quad & \mathbf{w}^\top \hat{\mu}_n = R \\ & \mathbf{w}^\top \mathbf{1} = 1 \end{aligned} \tag{SAA}$$

where

- $\hat{\Sigma}_{1:n}$ is the sample covariance matrix of $[\mathbf{x}_1, \ldots, \mathbf{x}_n]$.
- $\hat{\mu}_n$ is the sample average of the returns
- Solution: $\hat{\mathbf{w}}_{SAA}(R)$

In-sample vs. Out-of-sample performance

Three types of performance measures:

- In-sample performance: the performance of the learned action in the (training) sample, i.e. the data you used to learn
- Out-of-sample, or test, or generalization performance: the average performance of the learned action over all possible new observations
- Expected test, or true performance: the average performance of the learned action over all possible training sets and over all possible new observations

Note 1: for typical ML prediction problems, think error not performance. E.g. in-sample error, out-of-sample error, prediction error
Note 2: Training performance always overestimates (w.p. 1) both the out-of-sample and expected performances (why?)

In-sample vs. Out-of-sample return

In-sample (aka "training") return:

$$\hat{\mathbf{w}}_{SAA}^{\top}\hat{\boldsymbol{\mu}}_n$$

Out-of-sample (aka "test" or "generalization") return:

$$\mathbb{E}_{\mathbf{X}_{n+1}}[\hat{\mathbf{w}}_{SAA}^{\top}\mathbf{X}_{n+1}|\mathcal{X}_n] = \hat{\mathbf{w}}_{SAA}^{\top}\boldsymbol{\mu}$$

Expected test (aka "true") return:

$$\mathbb{E}_{\mathcal{X}_n}[\mathbb{E}_{\mathbf{X}_{n+1}}[\hat{\mathbf{w}}_{SAA}^{\top}\mathbf{X}_{n+1}|\mathcal{X}_n]]$$

In-sample vs. Out-of-sample risk

In-sample risk:

$$\hat{\mathbf{w}}_{SAA}^{\top} \hat{\Sigma}_{1:n} \hat{\mathbf{w}}_{SAA}$$

Out-of-sample risk:

$$Var_{\mathbf{X}_{n+1}}[\hat{\mathbf{w}}_{SAA}^{\top} \mathbf{X}_{n+1} | \mathcal{X}_n] = \hat{\mathbf{w}}_{SAA}^{\top} \Sigma \hat{\mathbf{w}}_{SAA}$$

Expected test risk:

$$\mathbb{E}_{\mathcal{X}_n}[Var_{\mathbf{X}_{n+1}}[\hat{\mathbf{w}}_{SAA}^{\top} \mathbf{X}_{n+1} | \mathcal{X}_n]]$$

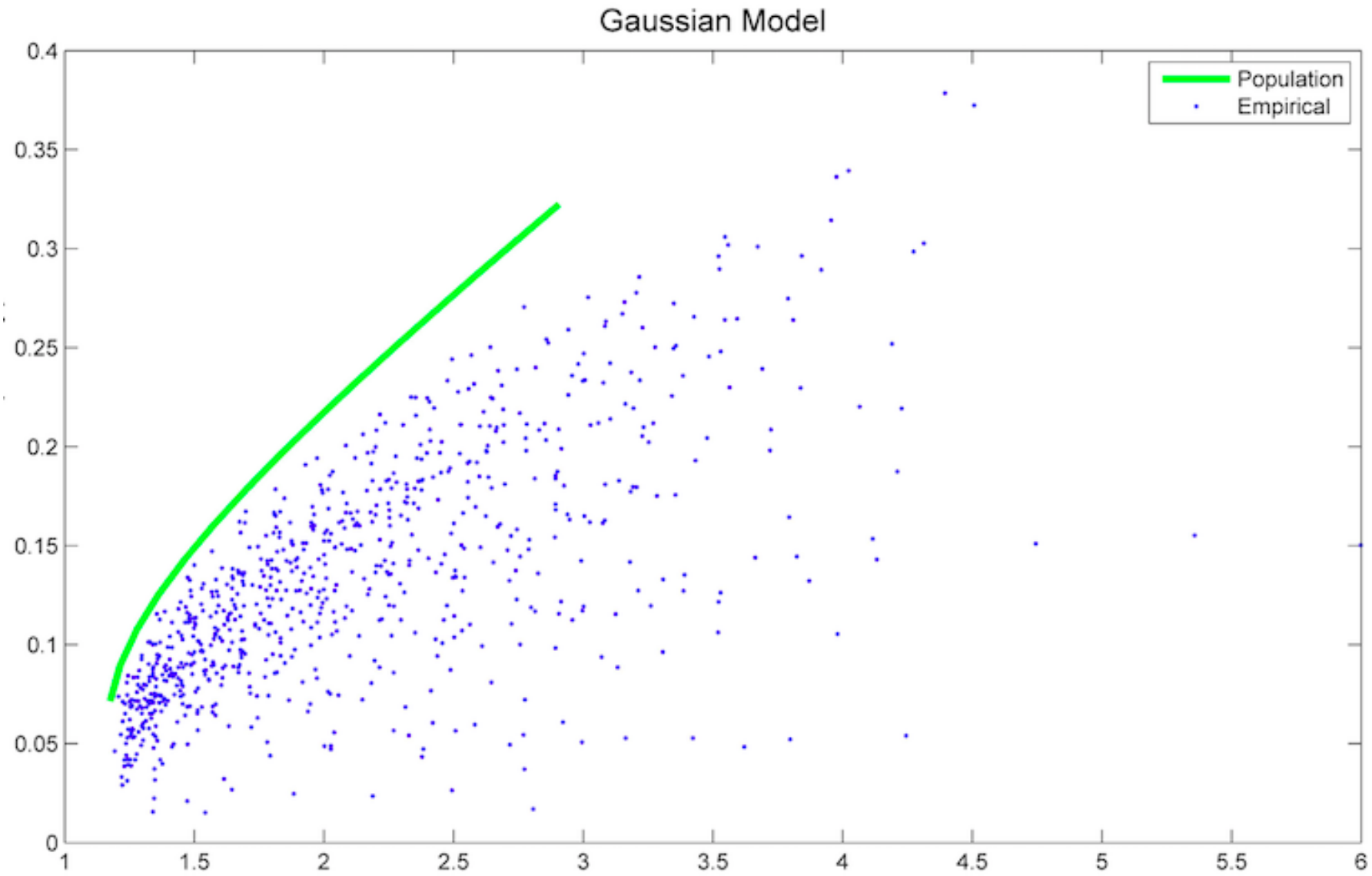
Performance of SAA: Simulated Data

Fix (ν, Q) and target return level R. Then for $b = 1, \ldots, B$,

- Generate $\mathcal{X}_{b,n} = [\mathbf{x}_{b,1}, \ldots, \mathbf{x}_{b,n}]$, where $\mathbf{X}_{b,i} \overset{iid}{\sim} \mathcal{N}(\nu, Q)$ for all $i = 1, \ldots, n$
- Solve the SAA problem for $\hat{\mathbf{w}}_{b,SAA}$
- Compute its out-of-sample return and risk: $\hat{\mathbf{w}}_{b,SAA}^{\top} \nu$ and $\hat{\mathbf{w}}_{b,SAA}^{\top} Q \, \hat{\mathbf{w}}_{b,SAA}$

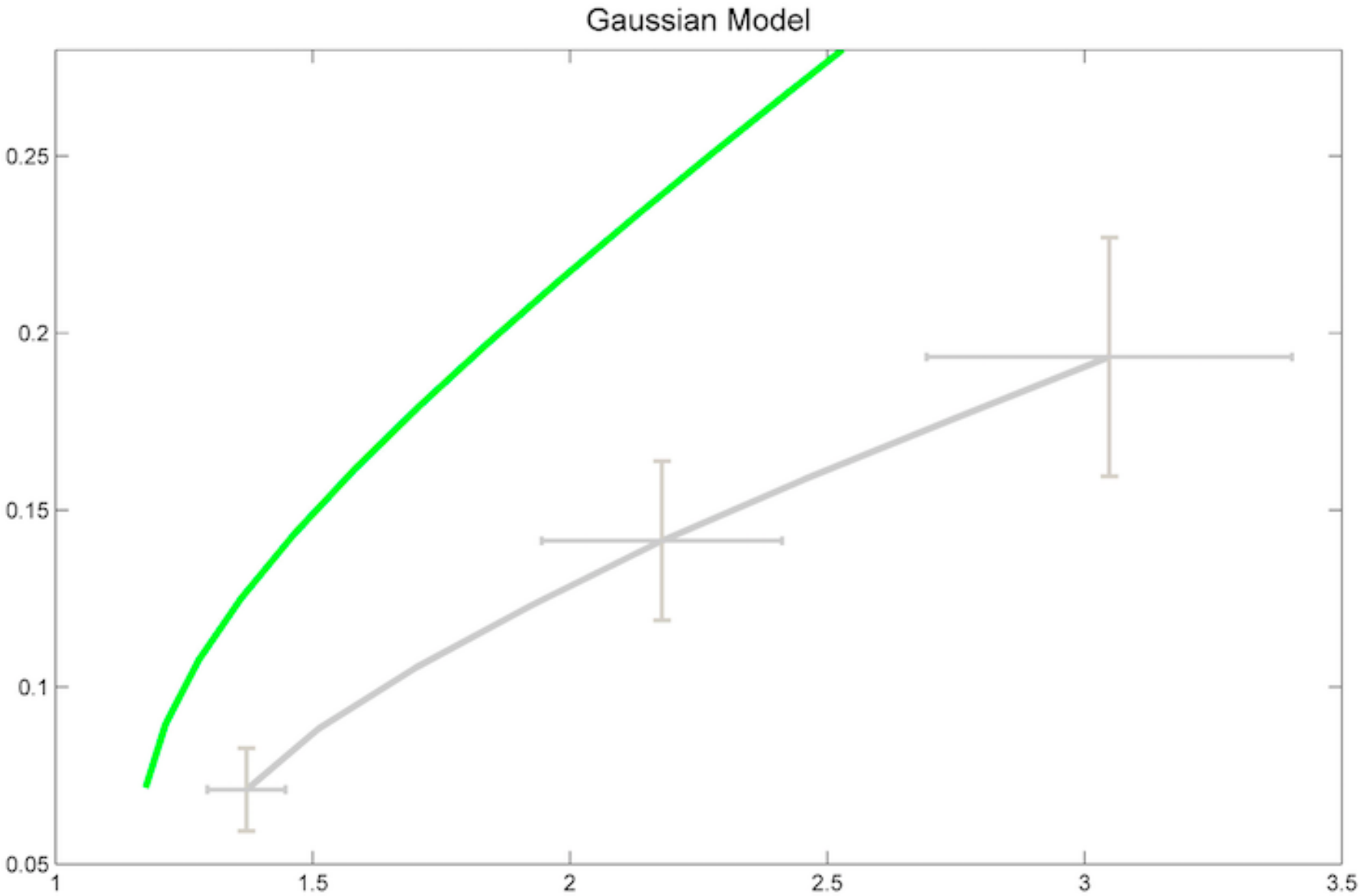
Performance of SAA

Return vs. Risk



Performance of SAA

Return vs. Risk



SAA is an error-maximizing algorithm

- Although SAA makes intuitive sense, it is highly unreliable for portfolio optimization with real stock return data
- This is well-documented across finance, statistics and OR:
  - Markowitz: Frankfurter et al. (1971), Frost & Savarino (1986, 1988b), Michaud (1989), Best & Grauer (1991), Chopra & Ziemba (1993), Broadie (1993), Lim et al. (2011)
- Michaud (1989): The (in-sample) portfolio optimization solution is an “error-maximizing” solution

Regularization

- Regularization: perturbing a linear operator problem for improved stability of solution [Ivanov (1962), Phillips (1962), Tikhonov (1963)]
- E.g. Least-squares regression with regularization:

$$\min_{\beta \in \mathbb{R}^p} ||\mathbf{y} - \mathbf{X}\beta||_2 + \lambda_n ||\beta||_k,$$

  where λ_n is the degree of regularization, and $k = 1$ (LASSO), $k = 2$ (ridge regression) yield popular penalty functions.
- Intuition: perturbing the in-sample problem reduces over-fitting; it adds bias but can improve the variance, which is good for generalization
- In general, $L-1$ norm penalty yields sparse (many elements are exactly zero) solution vector and $L-2$ norm penalty yields dense (many small, but non-zero elements) solution vector
- While these have justifications in regression problems, it's not clear why one would want sparse or dense portfolio solutions

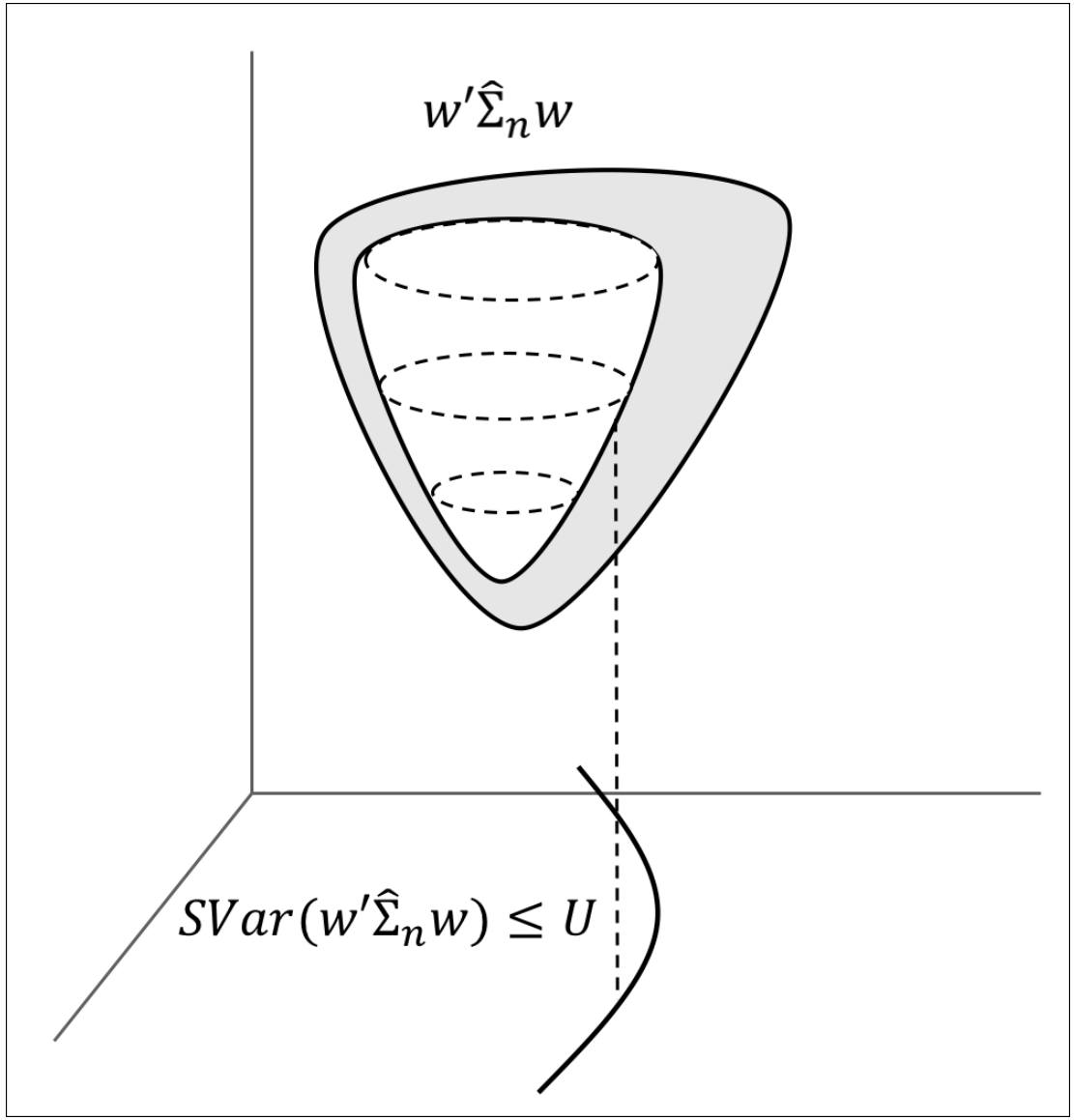
Performance-based regularization (PBR)

- Performance-based regularization: perturb portfolio problem for improved performance of the solution

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^\top \hat{\Sigma}_n \mathbf{w} \\ s.t. \quad & \mathbf{w}^\top \hat{\mu}_n = R \\ & \mathbf{w}^\top \mathbf{1} = 1 \\ & SVar(\mathbf{w}^\top \hat{\Sigma}_n \mathbf{w}) \leq U \end{aligned}$$

- Intuition: penalize solutions $\mathbf{w}$ associated with greater estimation errors of objective

Schematic for PBR



PBR for Mean-Variance problem

The sample variance of the sample variance of the portfolio, $SVar(w'\hat{\Sigma}_n w)$ is given by:

$$SVar(w'\hat{\Sigma}_n w) = \Sigma_{i=1}^{p}\Sigma_{j=1}^{p}\Sigma_{k=1}^{p}\Sigma_{l=1}^{p} w_i w_j w_k w_l \hat{Q}_{ijkl},$$

where

- $\hat{Q}_{ijkl} = \frac{1}{n}(\hat{\mu}_{4,ijkl} - \hat{\sigma}_{ij}^2\hat{\sigma}_{kl}^2) + \frac{1}{n(n-1)}(\hat{\sigma}_{ik}^2\hat{\sigma}_{jl}^2 + \hat{\sigma}_{il}^2\hat{\sigma}_{jk}^2)$,
- $\hat{\mu}_{4,ijkl}$ is the sample average estimator for $\mu_{4,ijkl}$, the fourth central moment of the elements of $\boldsymbol{X}$
- $\hat{\sigma}_{ij}^2$ is the sample average estimator for σ_{ij}^2, the covariance of the elements of $\boldsymbol{X}$.

PBR constraint for Markowitz is thus a quartic polynomial. However, determining whether a quartic function is convex or not is an NP-hard problem [Ahmadi et al. (2013)]

PBR for Mean-Variance problem

Convex approximation I

- Rank-1 approximation:

$$(\mathbf{w}^\top \hat{\alpha})^4 \approx \Sigma_{i=1}^{p} \Sigma_{j=1}^{p} \Sigma_{k=1}^{p} \Sigma_{l=1}^{p} w_i w_j w_k w_l \hat{Q}_{ijkl},$$

  where $\hat{\alpha}_i = \sqrt[4]{\hat{Q}_{iiii}}$.

- Approximate PBR constraint: $\mathbf{w}^\top \hat{\alpha} \leq \sqrt[4]{U}$

PBR for Mean-Variance problem

Convex approximation II

- Best convex quadratic approximation:

$$(\mathbf{w}^\top A\mathbf{w})^2 \approx \Sigma_{i=1}^{p}\Sigma_{j=1}^{p}\Sigma_{k=1}^{p}\Sigma_{l=1}^{p} w_i w_j w_k w_l \hat{Q}_{ijkl},$$

  such that the elements of A are as close as possible to the pair-wise terms of Q, i.e. $A_{ij}^2 \approx \hat{Q}_{ijij}$

- Solve semidefinite program: $A^* = \underset{A \succeq 0}{\operatorname{argmin}} ||A - Q_2||_F$, where Q_2 is a matrix with ij-th element equalling $\hat{Q}_{ijij}$ and $||\cdot||_F$ denotes the Frobenius norm:

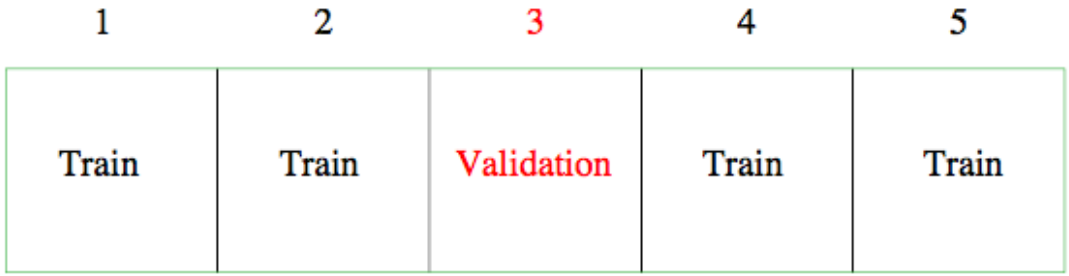
$$||A||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

- Approximate PBR constraint: $\mathbf{w}^\top A^*\mathbf{w} \leq \sqrt{U}$

Cross-Validation (CV)

Cross-Validation: if there's enough data, put aside some for tuning free parameters (the "validation data set"). E.g. 50% for training, 25% for validation and 25% for testing

k-fold Cross Validation: divide into $2 \leq k \leq n$ sub-training sets to maximize use of scarce training data.



Larger then k, the better the estimation of expected test error, but greater the computational burden and variance. $k = 5, 10$ are known to balance the trade-offs well. $k = n$ is leave-one-out CV

Performance-based CV

- CV: common technique in machine learning to tune free parameters
- k-fold CV: split training data into k equally-sized bins, train statistical model on every possible combination of $k-1$ bins, then tune parameter on the remaining bin.
- Performance-based k-fold CV: (1) search boundary for U_1 needs to be set carefully in order to avoid infeasibility and having no effect; (2) tune parameters by the Sharpe ratio, not by the mean squared error

Performance-based Cross-Validation

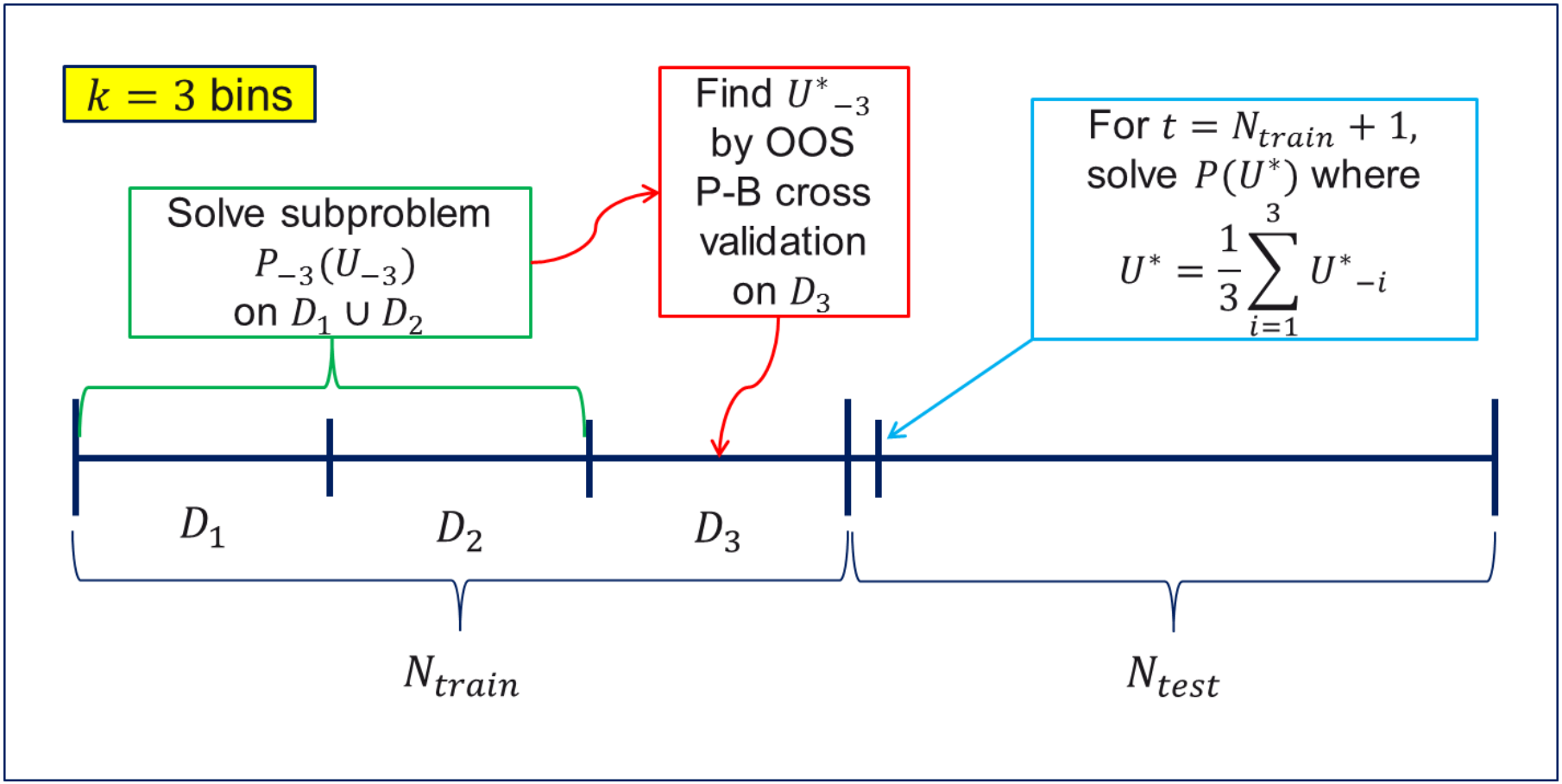


Figure: A schematic explaining the out-of-sample performance-based *k*-cross validation (OOS-PBCV) algorithm used to calibrate the constraint rhs, *U*, for the case $k = 3$. The training data set is split into *k* bins, and the optimal *U* for the entire training data set is found by averaging the best *U* found for each subset of the training data.

Empirical Results: Fama-French data sets

OOS Average Sharpe Ratio (Return/Std)

	FF 5 Industry p=5		FF 10 Industry p=10	
Mean-Variance R=0.04				
SAA	1.1459		1.1332	
	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	1.2603	1.3254	1.1868	1.2098
	(0.0411)	(0.0286)	(0.0643)	(0.0509)
PBR (PSD)	1.1836	1.1831	1.1543	1.1678
	(0.0743)	(0.071)	(0.0891)	(0.0816)
NS	1.0023		0.9968	
	(0.1404)		(0.1437)	
L1	1.0136	1.0386	1.1185	1.1175
	(0.1568)	(0.1396)	(0.1008)	(0.1017)
L2	0.9711	1.0268	1.0579	1.0699
	(0.1781)	(0.1452)	(0.1482)	(0.1280)

Parentheses: *p*-values of tests of differences from the SAA method.

Empirical Results: Fama-French data sets

OOS Average Sharpe Ratio (Return/Std)

	FF 5 Industry p=5		FF 10 Industry p=10	
Markowitz R=0.08				
SAA	1.1573		1.1225	
	2 bins	3 bins	2 bins	3 bins
PBR (rank-1)	1.3286	1.3551	1.1743	1.2018
	(0.0223)	(0.0208)	(0.0668)	(0.0510)
PBR (PSD)	1.1813	1.1952	1.1467	1.1575
	(0.0648)	(0.0614)	(0.0893)	(0.0844)
NS	0.9664		0.9405	
	(0.1514)		(0.1577)	
L1	0.9225	0.9965	1.0318	1.0779
	(0.1857)	(0.1403)	(0.1332)	(0.1181)
L2	0.9703	1.0284	1.0671	1.0776
	(0.1649)	(0.1398)	(0.1398)	(0.1209)

Parentheses: *p*-values of tests of differences from the SAA method.

Mean-CVaR Portfolio Optimization

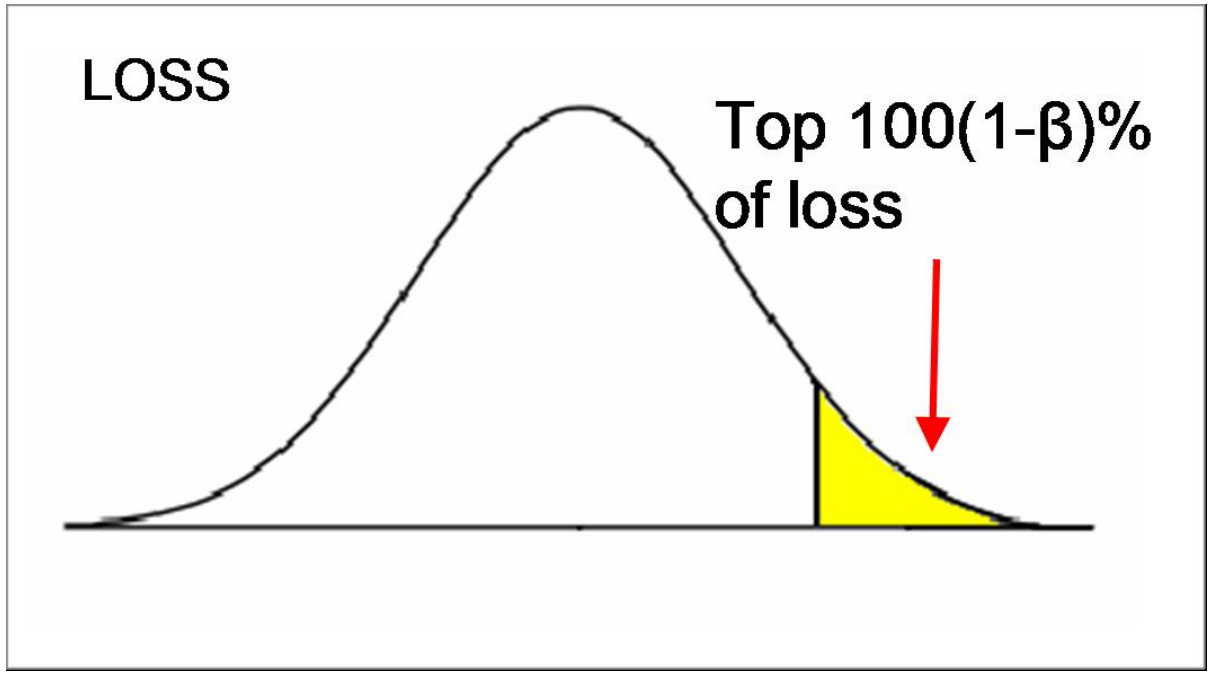
Consider the mean-Conditional Value-at-Risk portfolio optimization problem:

$$\begin{array}{lll} \min\limits_{\mathbf{w}} & \mathit{CVaR}(\mathbf{w}; \mathbf{X}, \beta) & \\ s.t. & \mathbf{w}^\top \boldsymbol{\mu} & = R \\ & \mathbf{w}^\top \mathbf{1} & = 1 \end{array} \tag{1}$$

where

- $\mathit{CVaR}(\mathbf{w}; \mathbf{X}, \beta) = \min\limits_{\alpha} \left\{ \alpha + \frac{1}{1-\beta} \mathbb{E}(-\alpha - \mathbf{w}^\top \mathbf{X}_i)^+ \right\}$

Conditional Value-at-Risk



- $CVaR(\mathbf{w}; \mathbf{X}, \beta) = \min_{\alpha} \left\{ \alpha + \frac{1}{1-\beta} \mathbb{E}(-\alpha - \mathbf{w}^\top \mathbf{X}_i)^+ \right\}$
- β = cutoff level, e.g. 95%, 99%
- Pros: tell you how thick the loss tail is; also a coherent risk measure [Acerbi & Tasche (2001)]

SAA for mean-CVaR problem

- Data: n iid observations of asset returns $\mathcal{X}_n = X_1, \ldots, X_n \sim P$

$$\begin{aligned} \min_{\mathbf{w}} \quad & \widehat{CVaR}_n(\mathbf{w}; \mathcal{X}_n, \beta) \\ s.t. \quad & \mathbf{w}^\top \hat{\boldsymbol{\mu}}_n = R \\ & \mathbf{w}^\top \mathbf{1} = 1 \end{aligned}$$

where

- $\hat{\boldsymbol{\mu}}_n$ is the sample average return;
- $\widehat{CVaR}_n(\mathbf{w}; \mathcal{X}_n, \beta) = \min_{\alpha}\{\alpha + \frac{1}{n(1-\beta)} \sum_{i=1}^{n} (-\alpha - \mathbf{w}^\top \mathbf{X}_i)^+\}$

PBR for mean-CVaR problem

Proposition

Suppose $\mathcal{X}_n = [X_1, \ldots, X_n] \overset{iid}{\sim} F$*, where* F *is absolutely continuous with twice continuously differentiable pdf. Then*

$$Var[\widehat{CVaR}_n(\mathbf{w}; \mathcal{X}_n, \beta)] = \frac{1}{n(1-\beta)^2} Var[(-\mathbf{w}^\top \mathcal{X}_n - \alpha_\beta(\mathbf{w}))^+] + O(n^{-2}),$$

where

$$\alpha_\beta(\mathbf{w}) = \inf\{\alpha : P(-\mathbf{w}^\top X \geq \alpha) \leq 1 - \beta\},$$

the Value-at-Risk (VaR) of the portfolio w *at level* β*.*

PBR for mean-CVaR problem

$$
\begin{aligned}
\min_{\mathbf{w}} \quad & \widehat{CVaR}_n(\mathbf{w}; \mathcal{X}_n, \beta) \\
s.t. \quad & \mathbf{w}^\top \hat{\mu}_n = R \\
& \mathbf{w}^\top \mathbf{1} = 1 \\
& \tfrac{1}{n(1-\beta)^2} z^\top \Omega_n z \leq U_1 \\
& \tfrac{1}{n} \mathbf{w}^\top \hat{\Sigma}_n \mathbf{w} \leq U_2 \\
& z_i = \max(0, -\mathbf{w}^\top X_i - \alpha), \quad i = 1, \ldots, n.
\end{aligned}
$$

- Not convex. Combinatorial optimization problem
- **Theorem**: convex relaxation, a QCQP, is tight
- Tune U_1 and U_2 via performance based k-fold CV

Empirical Results: mean-CVaR

OOS Average Sharpe Ratio (Return/CVaR)

	FF 5 Industry p=5		FF 10 Industry p=10	
Mean-CVaR R=0.04				
SAA	1.2137		1.0321	
	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	1.2113	1.1733	1.0506	1.1381
	(0.0554)	(0.0674)	(0.0638)	(0.0312)
PBR (mean only)	1.2089	1.1802	1.0994	1.0519
	(0.0746)	(0.0790)	(0.1051)	(0.1338)
PBR (both)	1.2439	1.2073	1.1112	1.1422
	(0.0513)	(0.0601)	(0.0691)	(0.0648)
L1	1.0112	1.0754	0.9254	0.9741
	(0.1497)	(0.1366)	(0.2293)	(0.1880)
L2	0.9650	1.0636	1.0031	0.9835
	(0.1780)	(0.1287)	(0.1512)	(0.1598)

Parentheses: *p*-values of tests of differences from the SAA method.

Empirical Results: mean-CVaR

OOS Average Sharpe Ratio (Return/CVaR)

	FF 5 Industry p=5		FF 10 Industry p=10	
Mean-CVaR R=0.08				
SAA	1.2487		1.0346	
	2 bins	3 bins	2 bins	3 bins
PBR (CVaR only)	1.2493	1.2098	1.0551	1.1433
	(0.0434)	(0.0462)	(0.0579)	(0.0323)
PBR (mean only)	1.2480	1.2088	1.0987	1.0470
	(0.0591)	(0.0693)	(0.1053)	(0.1384)
PBR (both)	1.2715	1.2198	1.1122	1.1449
	(0.0453)	(0.0544)	(0.0664)	(0.0639)
L1	0.8921	0.9836	0.9416	1.0087
	(0.1964)	(0.1572)	(0.2122)	(0.1645)
L2	0.9367	1.0801	1.0278	0.9947
	(0.1989)	(0.1179)	(0.1323)	(0.1530)

Parentheses: *p*-values of tests of differences from the SAA method.

Summary

- In general, in-sample optimal actions (predictions/decisions) do not generalize well out-of-sample. For the portfolio selection problem, solutions overweigh idiosyncratic observations in the training data.
- Regularization: L_1, L_2 norm penalties are standard, we explored more complex ones (PBR) to focus on the performance of a decision, rather than the prediction error.
- Performance-based Cross-Validation: data-driven methods to tune regularization parameters
- Can expect better out-of-sample performance with optimal amount of regularization that balances bias and variance.
- PBR solutions are better than SAA and L_1, L_2 regularized solutions (and other benchmarks) on well-known, publicly available data sets.

References

- Ban, Gah-Yi, Noureddine El Karoui, and Andrew EB Lim. "Machine Learning and Portfolio Optimization." *Management Science*, Articles in Advance, 21 Nov 2016.