Improved Shrinkage Prediction under a Spiked Covariance Structure

Gourab Mukherjee

University of Southern California

BIRS, April 9, 2019

Joint work with **Trambak Banerjee** and **Debashis Paul**

Shrinkage Prediction in Location Models with unknown Covariance

One sample Gaussian model:

$$\boxed{\texttt{Observed past:} \boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \;\; \texttt{Future:} \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0^{-1}\boldsymbol{\Sigma})}$$

- $\boldsymbol{\Sigma} \succ 0$ is unknown
- The past and the future are independent conditioned on $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

Goal: Based on observing $\boldsymbol{X}$ predict $\boldsymbol{Y}$ by $\hat{q}$ under an aggregative loss function $\mathcal{L}$ that is cumulative across co-ordinates. m_0: known.

Shrinkage Prediction in Location Models with unknown Covariance

One sample Gaussian model:

$$\boxed{\texttt{Observed past:}\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \;\; \texttt{Future:}\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0^{-1}\boldsymbol{\Sigma})}$$

- $\boldsymbol{\Sigma} \succ 0$ is unknown
- The past and the future are independent conditioned on $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

Goal: Based on observing $\boldsymbol{X}$ predict $\boldsymbol{Y}$ by $\hat{q}$ under an aggregative loss function $\mathcal{L}$ that is cumulative across co-ordinates. m_0: known.

Existing Literature. When Σ is known and:
(a) Homoskedastic: Extensive optimality studies on sperically symmetric estimators;
(b) Known Hetroskedasticity, Diagonal Σ: Xie, Kou, Brown' 12,16; Tran' 16, Weinstein, Ma, Brown, Zhang' 18, Sun et al, '18,;
(c) Known Correlated Structures, AR (1): Kong, Liu, Zhao, Zhou' 17.

Shrinkage Prediction in Location Models with unknown Covariance

One sample Gaussian model:

$$\boxed{\texttt{Observed past:} \boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \;\; \texttt{Future:} \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0^{-1}\boldsymbol{\Sigma})}$$

- $\boldsymbol{\Sigma} \succ 0$ is unknown
- The past and the future are independent conditioned on $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

Goal: Based on observing $\boldsymbol{X}$ predict $\boldsymbol{Y}$ by $\hat{q}$ under an aggregative loss function $\mathcal{L}$ that is cumulative across co-ordinates. m_0: known.

Existing Literature. When Σ is known and:
(a) Homoskedastic: Extensive optimality studies on sperically symmetric estimators;
(b) Known Hetroskedasticity, Diagonal Σ: Xie, Kou, Brown' 12,16; Tran' 16, Weinstein, Ma, Brown, Zhang' 18, Sun et al, '18,;
(c) Known Correlated Structures, AR (1): Kong, Liu, Zhao, Zhou' 17.

Side Information: Here, **we consider Σ is unknown** but we have side information in the form $\boldsymbol{W}_i$ that contain information on Σ but little information about θ. This side information can be essentially reduced:
$\boldsymbol{W}_i \overset{i.i.d.}{\sim} N_n(0, \Sigma), i = 1, \ldots, m \;\; \Leftrightarrow S \sim Wishart_n(m, \Sigma)$
Note: n dim, m side info. size

Shrinkage Prediction in Location Models with unknown Covariance

One sample Gaussian model:

$$\boxed{\texttt{Observed past:} \boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \;\; \texttt{Future:} \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0^{-1}\boldsymbol{\Sigma})}$$

- $\boldsymbol{\Sigma} \succ 0$ is unknown
- The past and the future are independent conditioned on $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

Goal: Based on observing $\boldsymbol{X}$ predict $\boldsymbol{Y}$ by $\hat{q}$ under an aggregative loss function $\mathcal{L}$ that is cumulative across co-ordinates. m_0: known.

Side Information: We have side information in the form $\boldsymbol{W}_i$ that contain information on Σ but little about θ.

Lagged data. Consider observing m vectors from a drift changing model across m time points: $W_t = \mu_t + \epsilon_t$ where $\epsilon_t \overset{i.i.d.}{\sim} N_n(0, \Sigma)$.
• Predicting W_C at Current time and the lag $C - m$ is huge, then, W_t will not be useful for the current location as it involves extrapolating too far.
• Assuming some regualrity in the drift process across time $\{\mu_t : 1 \leq t \leq m\}$ we can have $S := S(W_1, \ldots, W_m) \sim \text{Wishart}_n(\Sigma, \ \mathsf{df} \approx m,)$.

Spiked Covariance Structure

We assume a spiked covariance structure on the unknown $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \sum_{j=1}^{K} \ell_j \boldsymbol{p}_j \boldsymbol{p}_j' + \ell_0 (\boldsymbol{I} - \sum_{j=1}^{K} \boldsymbol{p}_j \boldsymbol{p}_j')$$

- $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K$ - orthonormal and $\ell_1 > \ldots > \ell_K > \ell_0 > 0$
- $K \ll n$ fixed but unknown

These kind of dependence structures arise in numerous applications that involve prediction in correlated models:

- Portfolio Selection [Karoui et al, 2013]
- Gene Expression Data-sets, [Fan et al., 2017]
- Health Care Management [Vahn et al, 2018]

*Note: In our framework can accomate the scenario $m, n \to \infty$ & $m/n \to 0$.

Shrinkage Prediction in Aggregative Models

Aggregative Model

Predicting a linear transformation of the unobserved future $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Y}$

Observed: $\boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \Sigma)$ Future: $\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0\,\Sigma)$

$$\boxed{\text{Our target is now linearly aggregated predictants: } \boldsymbol{V} = \boldsymbol{A}\boldsymbol{Y}}$$

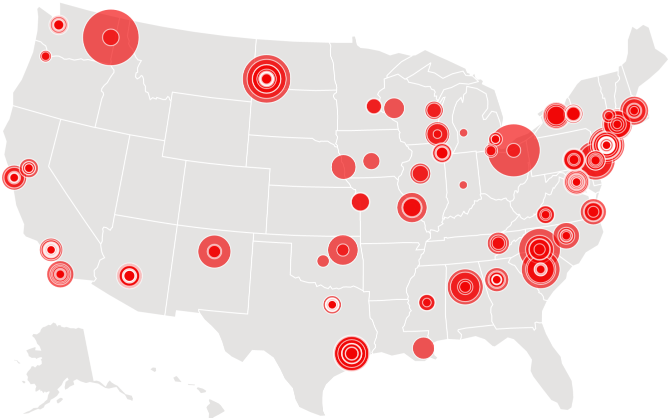
• The prediction problem is to make forecasts $\hat{\boldsymbol{q}} = \{\hat{q}_i(\boldsymbol{X}) : 1 \leq i \leq p\}$ based on the past data $\boldsymbol{X}$ such that $\hat{\boldsymbol{q}}$ optimally predicts $\boldsymbol{V}$.

• $\dim(A) = p \times n$ with $p \leq n$ and AA' is invertible.

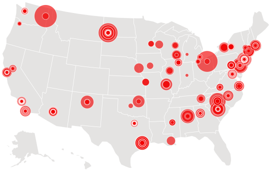
$$\boxed{\text{When } A = I_n \text{ we are back to the former disaggregate level model}}$$

Motivating Example - Inventory Management problem

Sale of Coffee in the week of Oct 31, 2011



Motivating Example - Inventory Management problem



Background - distributors and retailers

- based on past sales data, need to predict future demands across many stores.
- balance the trade-offs between **stocking too much** versus **stocking too little**.
- Incorporating co-dependencies in the demands among different stores is potentially useful.

Goal: predict demand for product $\mathcal{P}$ in week across n outlets.

- must leverage the co-dependencies in demands among the n stores.
- Forecasting future sales translates to a high-dimensional prediction problem.
- Aggregated problem - forecast sales aggregated across $p \leq n$ outlets.

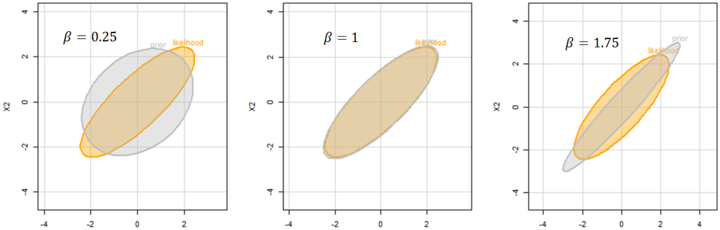
The co-dependencies between the demands is usually unknown.

A flexible conjugate Prior on θ (dis-aggregative model)

We impose a class of conjugate priors on the location parameter $\boldsymbol{\theta}$ that is related to the unknown covariance $\boldsymbol{\Sigma}$ by hyper-parameters β and τ:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \tau, \beta) \sim N_n\left(\underbrace{\boldsymbol{\eta}}_{location}, \underbrace{\tau \cdot \boldsymbol{\Sigma}^{\beta}}_{scale \times structure} \right)$$

- $\boldsymbol{\eta} \in \mathbb{R}^n$ and $\tau > 0$
- Power / Shape hyper-parameter: $\beta \geq 0$
- Non-exchangeability when $\beta > 0$
- Widely used in finance literature [Kozak et al (2017)]
  - $\beta = 0$: completely exchangeable
  - $\beta = 1$: same structure as the data
  - $\beta > 1$: prior more concentrated in dominant variability directions.



A flexible conjugate Prior on θ (dis-aggregative model)

We impose a class of conjugate priors on the location parameter $\boldsymbol{\theta}$ that is related to the unknown covariance $\boldsymbol{\Sigma}$ by hyper-parameters β and τ:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Sigma},\tau,\beta) \sim N_n\Big(\underbrace{\boldsymbol{\eta}}_{location}\, , \quad \underbrace{\tau \cdot \boldsymbol{\Sigma}^{\beta}}_{scale \times structure} \quad \Big)$$

- $\boldsymbol{\eta} \in \mathbb{R}^n$ and $\tau > 0$
- Power / Shape hyper-parameter: $\beta \geq 0$
- Non-exchangeability when $\beta > 0$
- Widely used in finance literature [Kozak et al (2017)]

In dis-aggregative model, the predictive distribution of $\boldsymbol{V}$ is given by:

$$N_n\big(\boldsymbol{\eta A 1} + G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \eta\boldsymbol{1}), G_{1,0,\beta} + m_0^{-1} G_{0,1,0}\big) \quad \text{where,}$$

$$G_{r,\alpha,\beta} = (\check{\boldsymbol{\Sigma}}_1^{-1} + \tau^{-1}\check{\boldsymbol{\Sigma}}_\beta^{-1})^{-r}\check{\boldsymbol{\Sigma}}_1^{\alpha} \quad \text{and} \quad \check{\boldsymbol{\Sigma}}_\beta = \boldsymbol{A}\boldsymbol{\Sigma}^{\beta}\boldsymbol{A}^T.$$

• As A does not always commute with Σ, in the aggregative model $\check{\boldsymbol{\Sigma}}_\beta^{-1}$ and $\check{\boldsymbol{\Sigma}}_1^{-1}$ have different eigen vectors unless $\beta = 1$. This increases the complexity in $G_{r,\alpha,\beta}$ due to aggregation.

Loss Functions

Recall $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Y}$ and let $\Lambda = (\boldsymbol{\theta}, \boldsymbol{\Sigma})$.

Loss associated with the i^{th} aggregator:

$$\mathcal{L}_i(\Lambda, \widehat{q}_i(\boldsymbol{A}, \boldsymbol{x})) = d_U(V_i - \widehat{q}_i)^+ + d_O(\widehat{q}_i - V_i)^+$$

d_U: under estimation loss d_O: over estimation loss

$$\text{Agglomerative Loss: } \mathcal{L}(\Lambda, \widehat{\boldsymbol{q}}) = \frac{1}{p}\sum_{i=1}^{p} \mathcal{L}_i(\Lambda, \widehat{q}_i(\boldsymbol{A}, \boldsymbol{x}))$$

Popular Loss Functions:

- **Symmetric Loss:** $d_U = d_O$
- **Asymmetric Loss:**
  - Quantile loss, $d_U / d_O = b \neq 1$
  - Linex loss, d_O is exponential and d_U is linear

Bayes Predictors

- $\check{\boldsymbol{\Sigma}}_\beta = \boldsymbol{A}\boldsymbol{\Sigma}^\beta \boldsymbol{A}^T$.
- $G_{r,\alpha,\beta} := G_{r,\alpha,\beta}(\boldsymbol{\Sigma}, \boldsymbol{A}) = (\check{\boldsymbol{\Sigma}}_1^{-1} + \tau^{-1}\check{\boldsymbol{\Sigma}}_\beta^{-1})^{-r}\check{\boldsymbol{\Sigma}}_1^\alpha$

If Σ were known, the Bayes predictor for $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{X}$ is

$$\boxed{\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{\Sigma}, \eta, \tau, \beta) = \eta \boldsymbol{e}_i^T \boldsymbol{A}\mathbf{1} + \boldsymbol{e}_i^T G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \eta\mathbf{1}) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)}$$

where, $\mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)$ is given by:

⋆ for generalized absolute loss where $d_U/d_O = b_i$ for the i th aggregator:

$$\Phi^{-1}(b_i)\Big(\boldsymbol{e}_i^T G_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T G_{0,1,0}\boldsymbol{e}_i\Big)^{1/2}$$

⋆ for linex loss with a_i being the asymmetry of the i th aggregator:

$$-\frac{a_i}{2}\Big(\boldsymbol{e}_i^T G_{1,0,\beta}\boldsymbol{e}_i + m_0^{-1}\boldsymbol{e}_i^T G_{0,1,0}\boldsymbol{e}_i\Big)$$

⋆ for symmetric quadratic loss: 0.

Bayes Predictors

- $\check{\boldsymbol{\Sigma}}_\beta = \boldsymbol{A}\boldsymbol{\Sigma}^\beta \boldsymbol{A}^T$.
- $G_{r,\alpha,\beta} := G_{r,\alpha,\beta}(\boldsymbol{\Sigma}, \boldsymbol{A}) = (\check{\boldsymbol{\Sigma}}_1^{-1} + \tau^{-1}\check{\boldsymbol{\Sigma}}_\beta^{-1})^{-r}\check{\boldsymbol{\Sigma}}_1^\alpha$

If Σ were known, the Bayes predictor for $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{X}$ is

$$\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{\Sigma}, \eta, \tau, \beta) = \eta \boldsymbol{e}_i^T \boldsymbol{A}\boldsymbol{1} + \boldsymbol{e}_i^T G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \eta\boldsymbol{1}) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)$$

Disaggregative vs Aggregative Models.

If $A = I$: in disaggregative model:

$$G_{r,\alpha,\beta} = H_{r,\alpha,\beta}, \text{ where } H_{r,\alpha,\beta}(\boldsymbol{\Sigma}) = (\boldsymbol{\Sigma}^{-1} + \tau^{-1}\boldsymbol{\Sigma}^{-\beta})^{-r}\boldsymbol{\Sigma}^\alpha$$

Note, that H involves Σ instead of $\check{\Sigma}_\beta$ and unlike aggregative models H has the same eigen vectors as Σ.

In aggregative models: $\tau^{-r} G_{r,\alpha,\beta}$ equals

$$\left\{\boldsymbol{A}H_{0,\beta,0}\boldsymbol{A}^T\left[\boldsymbol{A}\left(\tau H_{0,\beta,0} + H_{0,1,0}\right)\boldsymbol{A}^T\right]^{-1}\boldsymbol{A}H_{0,1,0}\boldsymbol{A}^T\right\}^r\left(\boldsymbol{A}H_{0,1,0}\boldsymbol{A}^T\right)^\alpha .$$

Evaluating Bayes Predictors under dependence

Recall: If Σ were known, the Bayes predictor for $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Y}$ is

$$\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{\Sigma}, \eta, \tau, \beta) = \eta \boldsymbol{e}_i^T \boldsymbol{A}\boldsymbol{1} + \boldsymbol{e}_i^T G_{1,-1,\beta} \boldsymbol{A}(\boldsymbol{X} - \eta \boldsymbol{1}) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)$$

• Thus, we need good estimates *based on* $\boldsymbol{X}$ *and* $\{W_i : 1 \le i \le m\}$ *only and without knowledge of* Σ of **quadratic forms** $b^T G_{r,\alpha,\beta} b$ involving $G_{r,\alpha,\beta}$.

• In Disaggregative model, estimating these quadratic forms involving G reduces to estimating quadratic forms involving H which is *comparatively* easier. We concentrate on estimating $b^T H_{r,\alpha,\beta}\, b$ first where $H_{r,\alpha,\beta}(\boldsymbol{\Sigma}) = (\boldsymbol{\Sigma}^{-1} + \tau^{-1}\boldsymbol{\Sigma}^{-\beta})^{-r}\boldsymbol{\Sigma}^{\alpha}$ and $||b||_2 = 1$.

Evaluating Bayes Predictors under dependence

Estimating $b^T H_{r,\alpha,\beta}\, b$: $H_{r,\alpha,\beta}(\mathbf{\Sigma}) = (\mathbf{\Sigma}^{-1} + \tau^{-1}\mathbf{\Sigma}^{-\beta})^{-r}\mathbf{\Sigma}^{\alpha}$, $||b||_2 = 1$.

Under spiked structure, efficient estimates of $\hat{\ell}_j$ of the eigen values and $\hat{p}_j$ of the K principal eigen vectors can be done. Consider:

$$\boxed{\hat{H}_{r,\alpha,\beta} = \sum_{j=1}^{K} \hat{\zeta}_j^{-2}\big(h_{r,\alpha,\beta}(\hat{\ell}_j) - h_{r,\alpha,\beta}(\hat{\ell}_0)\big)\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0)I}$$

where, $h_{r,\alpha,\beta}(x) = (x^{-1} + \tau^{-1}x^{-\beta})^{-r}x^{\alpha}$ is the scalar version of H and

$\zeta(x,\rho) = \Big[\dfrac{1-\rho/(x-1)^2}{1+\rho/(x-1)}\Big]^{1/2}$ and $\hat{\zeta}_j = \zeta(\hat{\ell}_j/\hat{\ell}_0, n/(m-1))$

$b^T\widehat{\boldsymbol{H}}_{\boldsymbol{r,\alpha,\beta}}b$ - bias corrected and consistent estimate of $b^T\boldsymbol{H}_{\boldsymbol{r,\alpha,\beta}}b$

- Asymptotic adjustments to the sample eigenvalues
- Phase transition phenomenon of the sample eigenvectors (Paul (2007))

Evaluating Bayes Predictors under dependence

Estimating $b^T H_{r,\alpha,\beta}\, b$: $H_{r,\alpha,\beta}(\mathbf{\Sigma}) = (\mathbf{\Sigma}^{-1} + \tau^{-1}\mathbf{\Sigma}^{-\beta})^{-r}\mathbf{\Sigma}^{\alpha}$, $||b||_2 = 1$.

Under spiked structure, efficient estimates of $\hat{\ell}_j$ of the eigen values and $\hat{p}_j$ of the K principal eigen vectors can be done. Consider:

$$\hat{H}_{r,\alpha,\beta} = \sum_{j=1}^{K} \hat{\zeta}_j^{-2}\big(h_{r,\alpha,\beta}(\hat{\ell}_j) - h_{r,\alpha,\beta}(\hat{\ell}_0)\big)\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0)I$$

Asymptotic consistency: $\mathbf{\Sigma}$ spike structure, $m/n > 0$ as $n \to \infty$

Uniformly over $\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0$ and $\boldsymbol{b} \in \mathcal{B}$ such that $|\mathcal{B}| = O(n^c)$ for any fixed $c > 0$ and $||\boldsymbol{b}||_2 = 1$, we have for all $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$

$$\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0,\boldsymbol{b}\in\mathcal{B}} \left|\boldsymbol{b}^T \widehat{H}_{r,\alpha,\beta}\boldsymbol{b} - \boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b}\right| = O_p\Big(\sqrt{\frac{\log n}{n}}\Big)$$

Evaluating Bayes Predictors under dependence

`Consider:` $\hat{H}_{r,\alpha,\beta} = \sum_{j=1}^{K} \hat{\zeta}_j^{-2}\big(h_{r,\alpha,\beta}(\hat{\ell}_j) - h_{r,\alpha,\beta}(\hat{\ell}_0)\big)\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0)I$

`Recall:` $\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{AX}|\boldsymbol{\Sigma},\eta,\tau,\beta) = \eta\boldsymbol{e}_i^T\mathbf{1} + \boldsymbol{e}_i^T H_{1,-1,\beta}(\boldsymbol{X}-\eta\mathbf{1}) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma},\tau,\beta)$

`Propose` $\hat{\boldsymbol{q}}^{\mathsf{step1}}_{(\mathsf{loss})}(\eta,\tau,\beta)$: Use $\hat{H}$ in place of H above.

Asymptotic consistency: $\boldsymbol{\Sigma}$ spike structure, $m/n > 0$ as $n \to \infty$

Uniformly over $\tau \in \boldsymbol{T}_0, \beta \in \boldsymbol{B}_0$ and $\boldsymbol{b} \in \mathcal{B}$ such that $|\mathcal{B}| = O(n^c)$ for any fixed $c > 0$ and $||\boldsymbol{b}||_2 = 1$, we have for all $(r,\alpha) \in \{-1,0,1\} \times \mathbb{R}$

$$\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0,\boldsymbol{b}\in\mathcal{B}} \left|\boldsymbol{b}^T\widehat{H}_{r,\alpha,\beta}\boldsymbol{b} - \boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b}\right| = O_p\Big(\sqrt{\frac{\log n}{n}}\Big)$$

Consequently, conditionally on $\boldsymbol{X}$,

$$\frac{\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0} ||\hat{q}^{\mathsf{step1}}(\boldsymbol{X}|\boldsymbol{S},\eta,\tau,\beta) - \boldsymbol{q}^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma})||_\infty}{||\boldsymbol{X}-\boldsymbol{\eta}||_2 \vee 1} = O_p\Big(\sqrt{\frac{\log n}{n}}\Big)$$

Proposed Prediction Rule - CASP

Key idea:

- construct efficient estimates of quadratic forms $\boldsymbol{a}^T \boldsymbol{H}_{\boldsymbol{r},\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{b}$
- **introduce coordinate-wise shrinkage policy to further reduce variability of** $\hat{q}^{\mathsf{step1}}$

CASP - **C**oordinate-wise **A**daptive **S**hrinkage **P**rediction

$$\hat{q}_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i^*) = \boldsymbol{e}_i^T \boldsymbol{A}\boldsymbol{\eta}_0 + \boldsymbol{f_i^*} \boldsymbol{e}_i^T \widehat{\boldsymbol{H}}_{\boldsymbol{1},\boldsymbol{-1},\boldsymbol{\beta}} \boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\eta}) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \tau, \beta)$$

- $b^T \widehat{\boldsymbol{H}}_{\boldsymbol{r},\boldsymbol{\alpha},\boldsymbol{\beta}} b$ - bias corrected and consistent estimate of $b^T \boldsymbol{H}_{\boldsymbol{r},\boldsymbol{\alpha},\boldsymbol{\beta}} b$
  - Phase transition phenomenon of the sample eigenvalues and eigenvectors
- $\boldsymbol{f_i^*}$ - coordinate wise shrinkage factor
  - Depends only on covariance level information through $\boldsymbol{W}$
  - Corresponds to actual reduction in marginal variability of q_i^{cs}
- This class of predictors includes our step1 predictor when $f_i = 1$ for all i. $\hat{q}_i^{\mathsf{step1}} = \hat{q}_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i = 1)$

Improving efficiency through co-ordinate wise shrinkage

- $\hat{q}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i = 1)$ - an asymptotically unbiased estimate of $\boldsymbol{q}^{\mathsf{Bayes}}$
- Average L_2 distance between them is non-trivial, however.

Recall $q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i^*) = \boldsymbol{e}_i^T \boldsymbol{\eta}_0 + \boldsymbol{f_i^*} \boldsymbol{e}_i^T \widehat{G}_{1,-1,\beta}(\boldsymbol{X} - \boldsymbol{\eta}) + \widehat{\mathcal{F}}_i^{\mathcal{L}_i}$

$$\text{Oracle choice: } \boxed{\boldsymbol{f_i^{\mathsf{OR}}} = \underset{f_i \in \mathbb{R}}{\arg\min}\, \mathbb{E}\Big\{\Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma})\Big)^2\Big\}}$$

- In general, $\boldsymbol{f_i^{\mathsf{OR}}} \in [0, 1]$
- Can be much smaller than 1 if the eigenvectors of $\boldsymbol{\Sigma}$ are relatively sparse
- $\hat{f}_i^*$ - a data driven choice such that $\sup_i |\hat{f}_i^* - \boldsymbol{f_i^{\mathsf{OR}}}| \to 0$ as $n \to \infty$

$$\hat{f}_i^* = \frac{\boldsymbol{e}_i^T \tau \hat{H}_{1,\beta-1,\beta} \boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{R} \boldsymbol{e}_i} \text{ where, } j(x) := x + \tau x^\beta,$$

$$\hat{R} = \tau \hat{H}_{1,\beta-1,\beta} + j(\hat{\ell}_0) \sum_{j=1}^{K} \hat{\zeta}_j^{-4} \Big(h_{1,-1,\beta}(\hat{\ell}_j) - h_{1,-1,\beta}(\hat{\ell}_0)\Big)^2 \hat{\boldsymbol{p}}_j \hat{\boldsymbol{p}}_j^T$$

Improving efficiency through co-ordinate wise shrinkage

Recall $q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i^*) = \boldsymbol{e}_i^T\boldsymbol{\eta}_0 + \boldsymbol{f_i^*}\boldsymbol{e}_i^T\widehat{G}_{1,-1,\beta}(\boldsymbol{X}-\boldsymbol{\eta}) + \widehat{\mathcal{F}}_i^{\mathcal{L}_i}$

$$\text{Oracle choice: } \boxed{\boldsymbol{f_i^{\mathsf{OR}}} = \arg\min_{f_i\in\mathbb{R}} \mathbb{E}\left\{\left(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma})\right)^2\right\}}$$

- In general, $\boldsymbol{f_i^{\mathsf{OR}}} \in [0,1]$
- Can be much smaller than 1 if the eigenvectors of $\boldsymbol{\Sigma}$ are relatively sparse
- $\hat{f}_i^*$ - a data driven choice such that $\sup_i|\hat{f}_i^* - \boldsymbol{f_i^{\mathsf{OR}}}| \to 0$ as $n\to\infty$

$$\hat{f}_i^* = \frac{\boldsymbol{e}_i^T\tau\hat{H}_{1,\beta-1,\beta}\boldsymbol{e}_i}{\boldsymbol{e}_i^T\hat{R}\boldsymbol{e}_i} \text{ where, } j(x) := x+\tau x^\beta,$$

$$\hat{R} = \tau\hat{H}_{1,\beta-1,\beta} + j(\hat{\ell}_0)\sum_{j=1}^K\hat{\zeta}_j^{-4}\left(h_{1,-1,\beta}(\hat{\ell}_j) - h_{1,-1,\beta}(\hat{\ell}_0)\right)^2\hat{\boldsymbol{p}}_j\hat{\boldsymbol{p}}_j^T$$

Oracle optimality of CASP: $\boldsymbol{\Sigma}$ spike structure, $m/n>0$ as $n\to\infty$

Conditionally on $\boldsymbol{X}$,

$$\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0}\frac{||\boldsymbol{q}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S},\hat{\boldsymbol{f}}^*) - \boldsymbol{q}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{f}^{\mathsf{OR}})||_2^2}{||\boldsymbol{q}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S},\boldsymbol{f}^{\mathsf{OR}}) - \boldsymbol{\eta}||_2^2} = O_p\left(\frac{\log n}{n}\right)$$

Evaluating Bayes Predictors in Aggregative Models

For a general $\boldsymbol{A}^{p\times n}$, $\tau^{-r}G_{r,\alpha,\beta}$ equals

$$\left\{\boldsymbol{A}H_{0,\beta,0}\boldsymbol{A}^T\left[\boldsymbol{A}\left(\tau H_{0,\beta,0}+H_{0,1,0}\right)\boldsymbol{A}^T\right]^{-1}\boldsymbol{A}H_{0,1,0}\boldsymbol{A}^T\right\}^r\left(\boldsymbol{A}H_{0,1,0}\boldsymbol{A}^T\right)^\alpha$$

- Substitute $\widehat{H}_{r,\alpha,\beta}$ in place of $H_{r,\alpha,\beta}$ in the above expression,

Asymptotic consistency: $\boldsymbol{\Sigma}$ spike structure, $m/n > 0$ as $n \to \infty$

Uniformly over $\tau \in \boldsymbol{T}_0$, $\beta \in \boldsymbol{B}_0$ and $\boldsymbol{b} \in \mathcal{B}$ such that $|\mathcal{B}| = O(n^c)$ for any fixed $c < 0$ and $||\boldsymbol{b}||_2 = 1$, we have for all $(r,\alpha) \in \{-1,0,1\} \times \mathbb{R}$

$$\sup_{\tau\in\boldsymbol{T}_0,\beta\in\boldsymbol{B}_0,\boldsymbol{b}\in\mathcal{B}}\left|\boldsymbol{b}^T\widehat{G}_{r,\alpha,\beta}\boldsymbol{b}-\boldsymbol{b}^T G_{r,\alpha,\beta}\boldsymbol{b}\right| = O_p\left(\max\left(\frac{\boldsymbol{p}}{\boldsymbol{n}},\ \sqrt{\frac{\log n}{n}}\right)\right)$$

- Consistency bounds deteriorate due to loss of commutativity for general $\boldsymbol{A}$ and the cost of its inversion is paid by the substitution rule for consistency
- Variance minimization via co-ordinate wise shrinkage can be done as before.

Real Data Illustration - Inventory Management

Background - distributors and retailers

- based on past sales data, need to predict future demands across many stores.
- balance the trade-offs between **stocking too much** versus **stocking too little**.
- Incorporating co-dependencies in the demands among different stores is potentially useful.

Data:

- Units of product $\mathcal{P}$ sold across $n \sim 1,200$ stores in week of Oct 31, 2011.
- Side information - Lagged data available for $m = 100$ weeks from December 31, 2007 to November 29, 2009.

Real Data - Loss Ratios

Table: Loss ratios across six predictive rules for four products.

Product	Method	K	Loss Ratio week w
Coffee (p = 31)	CASP	26	**0.999**
	Naïve	26	1.044
	Bcv	17	1.043
	POET	26	1.047
	Fact	26	1.009
	Unshrunk	-	1.838
Mayo (p = 30)	CASP	26	**0.995**
	Naïve	26	0.996
	Bcv	19	1.040
	POET	26	0.996
	Fact	26	0.999
	Unshrunk	-	1.084
Frozen Pizza (p = 33)	CASP	33	**0.998**
	Naïve	33	1.135
	Bcv	19	1.091
	POET	33	1.040
	Fact	33	1.020
	Unshrunk	-	6.701
Carb Beverages (p=33)	CASP	37	**0.984**
	Naïve	37	1.033
	Bcv	20	1.142
	POET	37	1.038
	Fact	37	1.059
	Unshrunk	-	8.885

Loss ratio for product $\mathcal{P}$:

$$\mathcal{L}_w(\boldsymbol{q}^{\mathsf{cs}}, \widehat{\boldsymbol{q}}) = \frac{\sum_{i=1}^{p} \left\{ b_i (V_i - \widehat{q}_i)^+ + h_i (\widehat{q}_i - V_i)^+ \right\}}{\sum_{i=1}^{p} \left\{ b_i (V_i - q_i^{\mathsf{cs}})^+ + h_i (q_i^{\mathsf{cs}} - V_i)^+ \right\}}$$

- $b_i = 0.95, \; h_i = 1 - b_i$
- $\boldsymbol{q}^{\mathsf{cs}}$ - CASP with $f_i = 1$
- $\widehat{\boldsymbol{q}}$ - any other predictive rule

- CASP: proposed method with data driven f_i
- Naive factor model without bias correction
- bi-cross-validation approach of Owen & Wang (2016)
- FactMLE algorithm of Khamaru & Mazumder (2018)

State-wise distribution of shrinkage factors

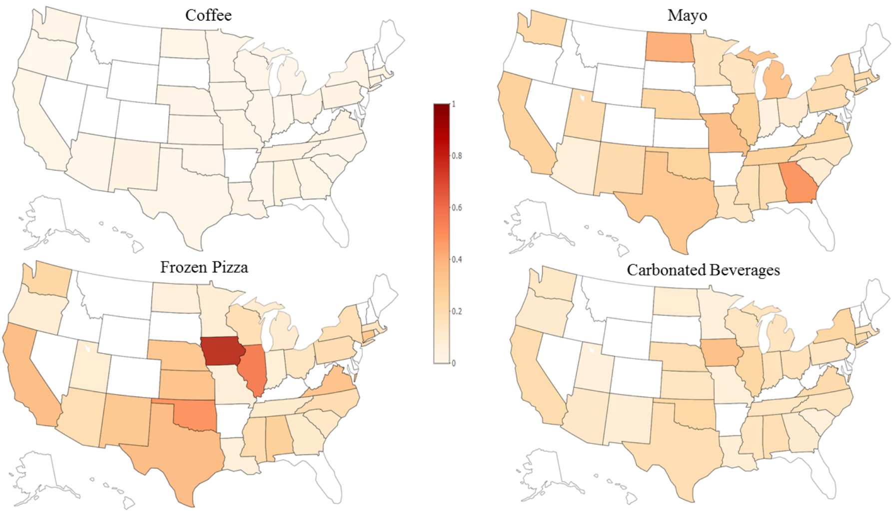


Figure: 1– the shrinkage factors of CASP by each state for the four products.

Closing Remarks

- We consider point prediction in location models with unknown covariance that has a spiked structure.
- A flexible non-exchangeable prior on the location parameter that depends on the unknown covariance is used.
- The prior induces skrinkage through the following hyper-parameters: (a) magnitude - that regulates amount of shrinkage (b) shape - that regulates the variability directions that are shrunken.
- We provide optimal evaluations of the Bayes predictors for a host of loss functions including symmetric and asymmetric losses. Bayes predictors involve functionals of unknown covariance.
- For such evaluations, we leverage the spiked covariance structure and use a simple substitution rule. Decision theoretic guarantees are provided for dis-aggregative as well as aggregative models.

THANKS!!

Manuscript available at: `http://www-bcf.usc.edu/ gourab/spiked.pdf`

R codes available at: `https://gmukherjee.github.io/Software/2018-08-15-casp/`

Supported in parts by: NSF DMS-1811866