Improved Shrinkage Prediction under a Spiked Covariance Structure

Trambak Banerjee
Analytics, Information and Operations Management
University of Kansas
Lawrence, KS 66045, USA

Gourab Mukherjee
Data Sciences and Operations
University of Southern California
Los Angeles, CA 90089, USA

Debashis Paul
Department of Statistics
University of California, Davis
Davis, CA 95616, USA

Editor:

Abstract

Keywords:

1. Introduction

2. Predictive model

Past observations and Future

$\bigg(\qquad \bigg)$

shape

scale

2.1 Aggregated prediction objectives

$\mathbb{R}$

2.2 Loss functions

Generalized absolute loss function

$\mathcal{L}$

Linex loss function

$$\mathcal{L} \qquad \left\{ \quad {}_{i} \quad {}_{i} \quad {}_{i} \qquad \right\}$$

2.3 Bayes predictors under known covariance

$\sum$ $\mathcal{L}$ *predictive loss*

$\mathbb{E}_{\boldsymbol{V}}$

predictive risk $\mathbb{E}_{\boldsymbol{X}}$

$\mathbb{E}_{\boldsymbol{AX}}$

$\int$

Lemma 1 **.** *Consider the hierarchical model in equations and . If were known, the unique minimizer of the integrated Bayes risk for coordinate is*

$$\mathcal{F}$$

where is the canonical basis vector with at the coordinate and . With , we have $\mathcal{F}$ *as follows:*

$$\mathcal{F} \qquad \begin{cases} \quad \Big(\qquad\qquad \Big) & \text{for generalized absolute loss} \\ -\Big(\qquad\qquad \Big) & \text{for Linex loss} \\ & \text{for quadratic loss} \end{cases}$$

2.4 Prediction under an unknown covariance and structural constraints

$\boldsymbol{j}$

$\mathbf{0}$

c c c c

c

c

$\mathbb{R}$

$\mathbf{0}$ $\boldsymbol{\mathcal{C}}$ 0 0

$\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$

Wishart

$\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$

$\mathbf{1}$ $\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$ $\boldsymbol{\mathcal{C}}$

predictive risk $\mathbb{E}_{\boldsymbol{AX}\ \boldsymbol{S}}$

$\mathbb{E}_{\boldsymbol{AX}\ \boldsymbol{S}}$

3. Proposed methodology for disaggregated model

3.1 Estimation of quadratic forms associated with Bayes predictors

A1 Asymptotic regime :

A2 Signi cant spike:

A1 **A2**

$$\sum \quad \Big(\qquad \Big)$$

$$\sum -$$

$$\sum \Big[- \qquad \Big(\quad - \Big) \qquad \Big] \qquad \qquad \Big(\quad \sum \quad \Big)$$

A3

A3

$\mathcal{B}$ $\mathbb{S}$

$\mathcal{B}$

Theorem 1A . *Under assumptions **A1**, **A2**, and **A3**, uniformly over , and $\mathcal{B}$ with , we have, for all* $\mathbb{R}$,

$$_{\boldsymbol{T}_0 \quad \boldsymbol{B}_0\, \boldsymbol{b} \;\; \mathcal{B}} \Big| \qquad \qquad \Big| \qquad \Big(\sqrt{\quad}\Big)$$

where the dependence of on has been kept implicit for notational ease.

$$\sum \qquad \qquad \sum$$

Lemma 2 . *Under assumptions **A1**, **A2** and **A3**, for any with we have,*

.

De nition 1 **.** *Under the hierarchical model of equations and , the proposed predictive rule for the disaggregated model is given by which is defined as*

$$\mathcal{F}$$

where

$$\mathcal{F} \quad \begin{cases} \Big(\qquad \Big) & \text{for generalized absolute loss} \\ -\Big(\qquad \Big) & \text{for Linex loss} \\ & \text{for quadratic loss} \end{cases}$$

and .

$$+ \qquad + \qquad - \qquad -$$

Lemma 3. *Under assumptions **A1**, **A2** and **A3**, uniformly over , , for all* $\mathbb{R}$*, we have, conditionally on ,*

$$\frac{\boldsymbol{T}_0 \quad \boldsymbol{B}_0 \Big\| \qquad \Big\|}{\Big\| \qquad \Big\|} \qquad \Big(\sqrt{\frac{}{}}\Big)$$

3.2 Improved predictive e ciency by coordinate-wise shrinkage

De nition 2 **.** *Consider a class of coordinate-wise shrinkage predictive rules* $\mathcal{Q}$ $\mathbb{R}$ *where*

$$\mathcal{F}$$

with $\mathcal{F}$ *as defined in Definition 1 and* $\mathbb{R}$ *is a shrinkage factor depending only on .*

$\mathcal{Q}$

Lemma 4. *Suppose that assumptions* ***A1***, ***A2*** *and* ***A3*** *hold. Under the hierarchical model of equations and , as ,*

(a) $\mathbb{E}\Big\{\Big(\quad \Big) \Big\}$ *is minimized at*

$$\frac{}{\mathcal{J}} \quad \Big(\sqrt{\frac{}{}}\Big)$$

where $\mathcal{J}$ *,* $\mathcal{J}$ *and the expectation is taken with respect to the marginal distribution of with fixed.*

(b) *For any fixed , , with probability 1,*

Moreover, let $\mathcal{M}$ *, where denotes the -dimensional projection matrix associated with the spiked eigenvalues of . Then, with as the scalar version of* $\mathcal{J}$ *, we have*

$$\mathcal{M} \quad \mathcal{M} \frac{}{\{ \quad \}} \quad \Bigg(\sqrt{\frac{}{}}\Bigg)$$

so that the leading term on the right hand side is less than 1.

(c) *Also, for any fixed and , we have with probability 1:*

$$\frac{\mathbb{E}}{\mathbb{E}}$$

where the expectations are taken with respect to the marginal distribution of with fixed.

$\mathcal{Q}$

$\mathcal{J}$

De nition 3 . *The coordinate-wise adaptive shrinkage prediction rule is given by* $\mathcal{Q}$ *with* *where*

and

$$\sum \left\{ \qquad \right\}$$

with *as the scalar version of* $\mathcal{J}$.

Lemma 5. *Under the hierarchical model of equations and ,*

$$\left(\sqrt{}\right)$$

$\mathcal{Q}$

$\mathcal{Q}$

Theorem 2A . *Under assumptions **A1**, **A2** and **A3**, and the hierarchical model of equations and , we have, conditionally on ,*

$$\boldsymbol{T}_0 \quad \boldsymbol{B}_0$$

3.3 Calibration of the tuning parameters

$\mathcal{J}$ $\mathcal{J}$

$\mathcal{J}$ $\mathcal{J}$

$\mathcal{J}$ $\mathcal{J}$

$\boldsymbol{T}_0$ $\boldsymbol{B}_0$ $\mathcal{J}$ ( ) $\mathcal{J}$ ( )

4. Methodology for the aggregated model

[( )]

[( )]

A4 Aggregation matrix: $\mathbb{R}$

De nition 4 **.** *For any fixed obeying assumption* ***A4****, consider a class of coordinate-wise shrinkage predictive rules* $\mathcal{Q}_{\boldsymbol{A}}$
$\mathbb{R}$ *where*

$\mathcal{F}$

and $\mathcal{F}$ *are the estimates of* $\mathcal{F}$ *as defined in Lemma 1 with replaced by and* $\mathbb{R}$ *are shrinkage factors depending only on and . The* *coordinate-wise adaptive shrinkage predictive rule for the aggregated model is given by* $\mathcal{Q}_{\boldsymbol{A}}$ *with* *where*

$$\frac{\mathcal{N}}{\quad}$$

and

$$\mathcal{N}$$

$$\mathcal{N} \quad \sum \quad \Big(\qquad \Big)$$

with *as the scalar version of* $\mathcal{J}$.

$$-$$

Theorem 1B . *Under assumptions* ***A1****,* ***A2****,* ***A3*** *and* ***A4****, uniformly over* , *and* $\mathcal{B}$ *with* , *we have for all* $\mathbb{R}$

$$_{\boldsymbol{T}_0 \quad \boldsymbol{B}_0\, \boldsymbol{b} \;\; \mathcal{B}} \Big| \qquad \Big| \quad \Big\{ \quad \Big(- \sqrt{\frac{}{}} \Big) \Big\}$$

where the dependence of *on* *has been kept implicit for notational ease.*

Theorem 2B . *Under assumptions* ***A1****,* ***A2****,* ***A3*** *and* ***A4****, and the hierarchical model of equations* *and* *, we have, conditionally on* ,

$$_{\boldsymbol{T}_0 \quad \boldsymbol{B}_0} \frac{\big\| \qquad \big\|}{\big\| \qquad \big\|} \quad \Big\{ \quad \Big((-) \quad \frac{}{} \Big) \Big\}$$

$\mathcal{Q}_{\boldsymbol{A}}$

Remark 1. On the uncertainty in estimating -

—

Remark 2. Implementation and R package `casp` - `casp`

https://github.com/trambakbanerjee/casp

5. Simulation studies

esaBcv

FACTMLE

POET

https://github.com/trambakbanerjee/CASP_paper

5.1 Experiment 1

0

-

-

0

POET

Table 1: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ for Scenarios 1 and 2 under Experiment 1. The numbers in parenthesis are standard errors over 500 repetitions.

	Scenario 1: $(K, \tau, \beta) = (10, 0.5, 0.25)$				Scenario 2: $(K, \tau, \beta) = (10, 1, 1.75)$			
	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)
CASP	7 (0.04)	0.59 (0.002)	0.27 (0.002)	**0.95**	4 (0.04)	0.97 (0.004)	1.79 (0.006)	**1.00**
Bcv	3 (0.08)	0.58 (0.003)	0.26 (0.003)	1.14	1 (0.04)	1.00 (0.001)	1.75 (0.006)	4.24
FactMLE	7 (0.04)	0.57 ($< 10^{-3}$)	0.19 (0.001)	1.68	4 (0.04)	0.98 ($< 10^{-3}$)	1.55 (0.001)	4.58
POET	7 (0.04)	0.57 ($< 10^{-3}$)	0.18 ($< 10^{-3}$)	2.14	4 (0.04)	0.97 ($< 10^{-3}$)	1.53 ($< 10^{-3}$)	7.26
Naive	7 (0.04)	0.60 ($< 10^{-3}$)	0.24 (0.001)	1.36	4 (0.04)	1.00 ($< 10^{-3}$)	1.63 (0.004)	1.87

under scenario 2 wherein the REE of CASP is 1. This is not unexpected because with a fixed $\tau > 0$ and β growing above 1, the factor $\sum_{j=1}^{K} \hat{\zeta}_j^{-4}\left(h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}})\right)^2$ in the denominator of $\hat{f}_i^{\mathsf{prop}}$ becomes smaller in comparison to the numerator $\hat{\mathcal{N}}$ in Definition 4 and the improvement due to coordinate-wise shrinkage dissipates. From table 1, we see

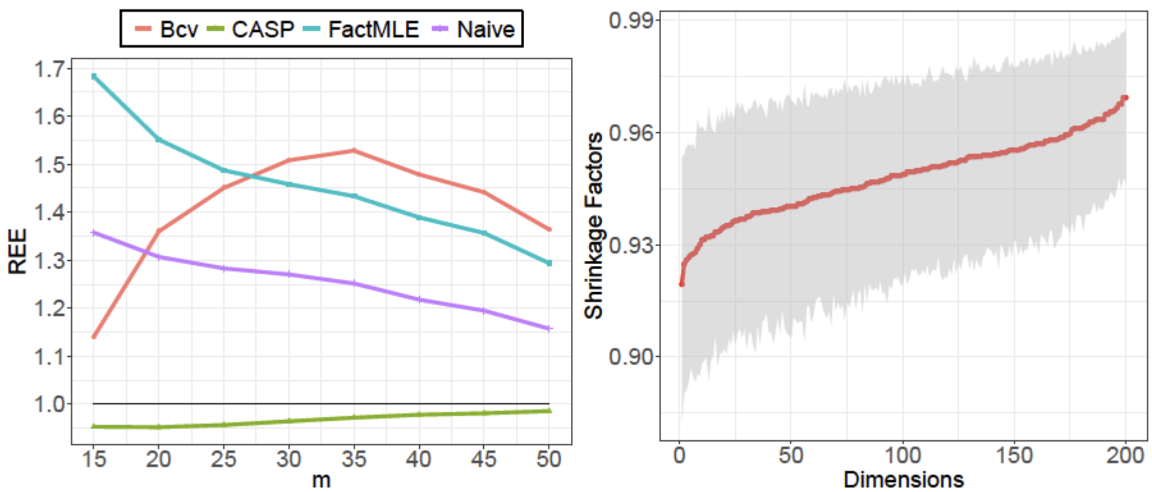


Figure 1: Experiment 1 Scenario 1 (Generalized absolute loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

that $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ is the most competitive predictive rule next to $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ across both the scenarios however it seems to suffer from the issue of under estimation of the number of factors K. We notice this behavior of $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ across all our numerical and real data examples.

The other three predictive rules, $\hat{\boldsymbol{q}}^{\mathsf{Fact}}$, $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ and $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$, exhibit poorer risk performances and this is not entirely surprising in this setting primarily because the four competing predictive rules considered here do not involve any asymptotic corrections to the sample eigenvalues and their eigenvectors whereas CASP uses the phase transition phenomenon of the sample eigenvalues and their eigenvectors to constructs consistent estimators of smooth functions of $\boldsymbol{\Sigma}$ that appear in the form of the Bayes predictive rules.

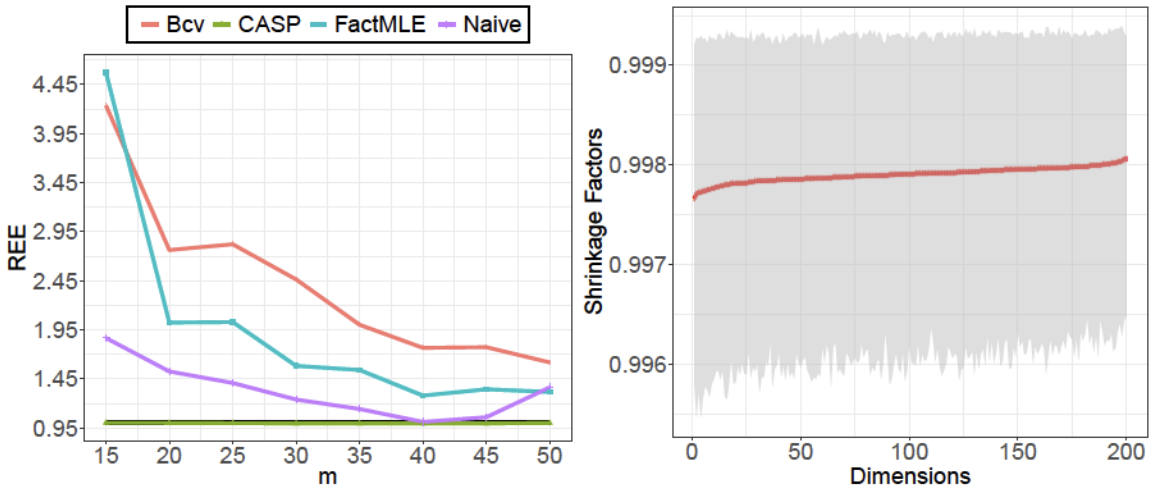


Figure 2: Experiment 1 Scenario 2 (Linex loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

From figure 1, we see that as m increases, $\hat{\boldsymbol{q}}^{\text{Naive}}$ performs better than $\hat{\boldsymbol{q}}^{\text{Bcv}}$. The spiked covariance structure considered in scenario 1 is substantially strong as there are $K = 10$ equispaced spikes between 20 and 80. The $\hat{\boldsymbol{q}}^{\text{Bcv}}$ method, which underestimates K more severely compared to $\hat{\boldsymbol{q}}^{\text{Naive}}$, performs worse as m increases for there is more information to estimate the spiked structure. The same phenomenon happens in scenario 2 where $\beta > 1$. However, when $\beta > 1$, most of the coordinate-wise shrinkage factors are close to 1 (see the right plot of figure 2) and so, the difference between CASP and $\hat{\boldsymbol{q}}^{\text{Naive}}$ is not much due to coordinate-wise shrinkage but mostly due to the biased estimation of the eigenvalues by the naive method.

5.2 Experiment 2

For experiment 2 we consider the setup of a static factor model with heteroscedastic noise and simulate our data according to the following model:

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{\theta} + \boldsymbol{B}\boldsymbol{\Gamma}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\Gamma}_t &\sim N_K(\boldsymbol{0}, \boldsymbol{I}_K) \\ \boldsymbol{\theta} &\sim N_n(\boldsymbol{\eta}_0, \tau\boldsymbol{\Sigma}^{\beta}) \text{ and } \boldsymbol{\epsilon}_t \sim N_n(\boldsymbol{0}, \boldsymbol{\Delta}_n), \end{aligned}$$

where $K \ll n$ represents the number of latent factors, $\boldsymbol{B}$ is the $n \times K$ matrix of factor loadings, $\boldsymbol{\Gamma}_t$ is the $K \times 1$ vector of latent factors independent of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\Delta}_n$ is an $n \times n$ diagonal matrix of heteroscedastic noise variances. In this model $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{\Delta}_n$ and coincides with the heteroscedastic factor models considered in Owen and Wang (2016); Fan et al. (2013); Khamaru and Mazumder (2019) for estimating $\boldsymbol{\Sigma}$. Thus the three competing predictive rules $\hat{\boldsymbol{q}}^{\text{Bcv}}$, $\hat{\boldsymbol{q}}^{\text{Poet}}$ and $\hat{\boldsymbol{q}}^{\text{Fact}}$ are well suited for prediction in this model. Factor models of this form are often considered in portfolio risk estimation (see for example Fan

0

0

0

-
-

0

0.87	**1.00**

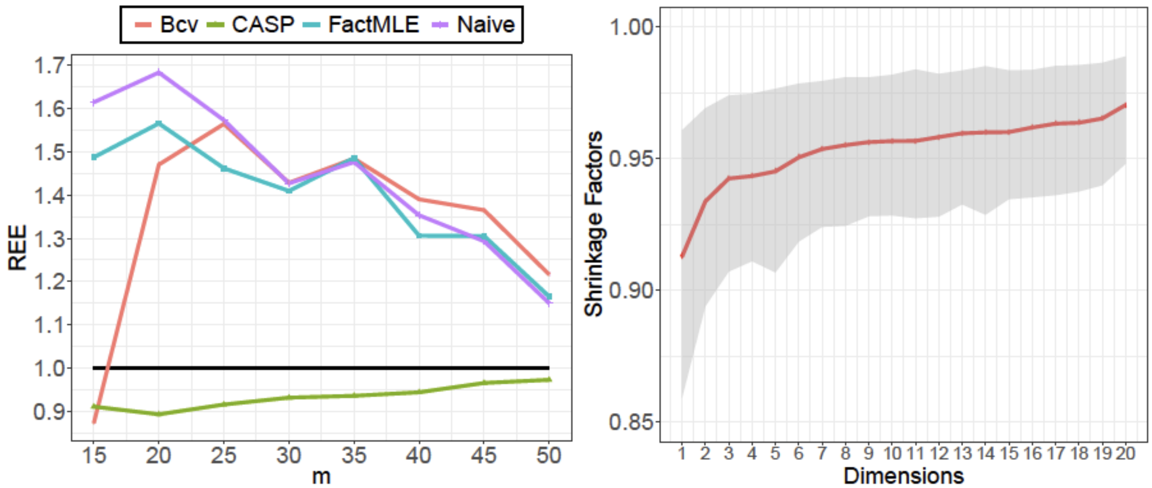


Figure 3: Experiment 2 Scenario 1 (Linex loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

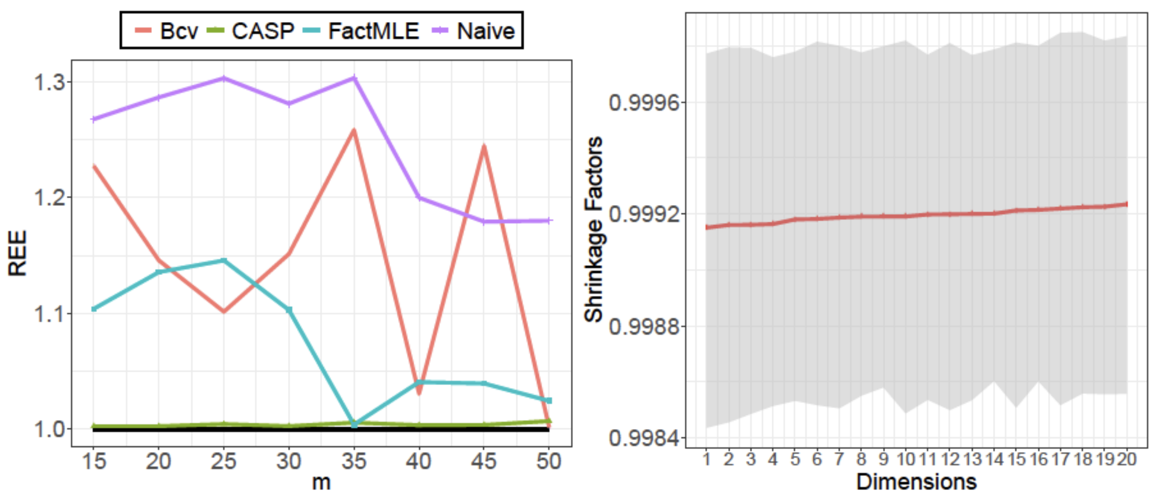


Figure 4: Experiment 2 Scenario 2 (Linex loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

5.3 Experiment 3

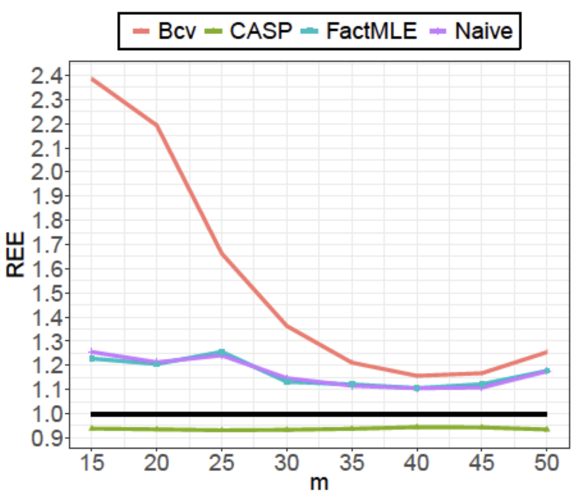
For experiment 3, we consider a slightly different setup where we do not impose a spike covariance structure on $\boldsymbol{\Sigma}$. Instead, we assume that $(\boldsymbol{\Sigma})_{ij} = Cov(X_i, X_j) = 0.9^{|i-j|}$ where $i, j = 1, \ldots, n$, thus imposing an AR(1) structure between the n coordinates of $\boldsymbol{X}$. As in experiment 1, we sample $\boldsymbol{\theta}$ from an $n = 200$ variate Gaussian distribution with mean vector

$\boldsymbol{\eta}_0 = \mathbf{0}$ and covariance $\tau\boldsymbol{\Sigma}^{\beta}$. We vary (τ, β) across two scenarios where we take (τ, β) as $(1, 0.5)$ and $(0.5, 2)$ in scenarios 1 and 2 respectively. We estimate $\boldsymbol{S}$ using the approach described in experiments 1 and 2, and sample $m_x = 1$ copy of $\boldsymbol{X}$ from $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with a goal to predict $\boldsymbol{AY}$ under a generalized absolute loss function with $h_i = 1 - b_i$ and b_i sampled uniformly from $(0.9, 0.95)$ for $i = 1, \cdots, p$. Here $\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ is independent of $\boldsymbol{X}$ and $\boldsymbol{A}$ is a fixed $p \times n$ sparse matrix with the $p = 20$ rows sampled independently from a mixture distribution with density $0.9\delta_0 + 0.1\,\mathit{Unif}(0, 1)$ and normalized to 1 thereafter. This sampling scheme is repeated over 500 repetitions and the REE of the competing predictive rules and CASP is presented in figures 5, 6 and table 3.

Table 3: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ for Scenarios 1 and 2 under Experiment 3. The numbers in parenthesis are standard errors over 500 repetitions.

	Scenario 1: $(\tau, \beta) = (1, 0.5)$				Scenario 2: $(\tau, \beta) = (0.5, 2)$			
	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)	$\hat{K}$	$\hat{\tau}$	$\hat{\beta}$	REE($\hat{\boldsymbol{q}}$)
CASP	7 (0.06)	1.09 (0.007)	0.40 (0.004)	**0.94**	7 (0.06)	0.57 ($< 10^{-3}$)	1.74 (0.001)	**1.00**
Bcv	1 (0.08)	0.95 (0.012)	0.36 (0.002)	2.39	1 (0.08)	0.59 ($< 10^{-3}$)	1.79 (0.002)	4.27
FactMLE	7 (0.06)	1.16 (0.001)	0.33 ($< 10^{-3}$)	1.23	7 (0.06)	0.57 ($< 10^{-3}$)	1.73 ($< 10^{-3}$)	1.08
POET	7 (0.06)	1.17 ($< 10^{-3}$)	0.33 ($< 10^{-3}$)	1.49	7 (0.06)	0.57 ($< 10^{-3}$)	1.73 ($< 10^{-3}$)	1.27
Naive	7 (0.06)	1.16 (0.001)	0.34 (0.001)	1.26	7 (0.06)	0.57 ($< 10^{-3}$)	1.73 ($< 10^{-3}$)	1.11

In this setup, the departure from the factor model leads to a poorer estimate of β for CASP than what was observed under experiments 1 and 2, however, the REE of CASP continues to be the smallest amongst all the other competing rules. When $\beta = 2$ (scenario 2), $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ and $\hat{\boldsymbol{q}}^{\mathsf{S}}$ are almost identical in their performance. Amongst the competing methods



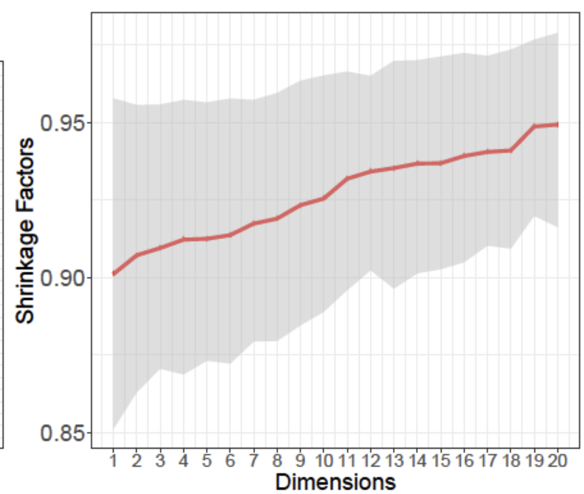


Figure 5: Experiment 3 Scenario 1 (Generalized absolute loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

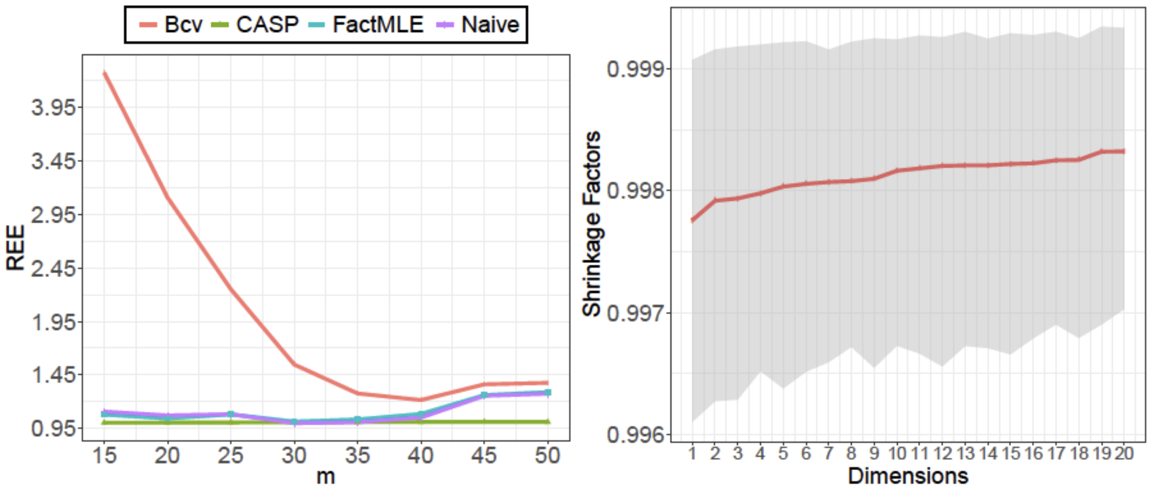


Figure 6: Experiment 3 Scenario 2 (Generalized absolute loss): Left - Relative Error estimates as m varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over 500 repetitions at $m = 15$ and sandwiched between its 10^{th} and 90^{th} percentiles

$\hat{q}^{\mathsf{Bcv}}$ has the highest REE, possibly exacerbated by the departure from a factor based model considered in this experiment whereas this seems to have a comparatively lesser impact on CASP indicating potential robustness of CASP to misspecifications of the factor model.

6. Real data illustration with groceries sales data

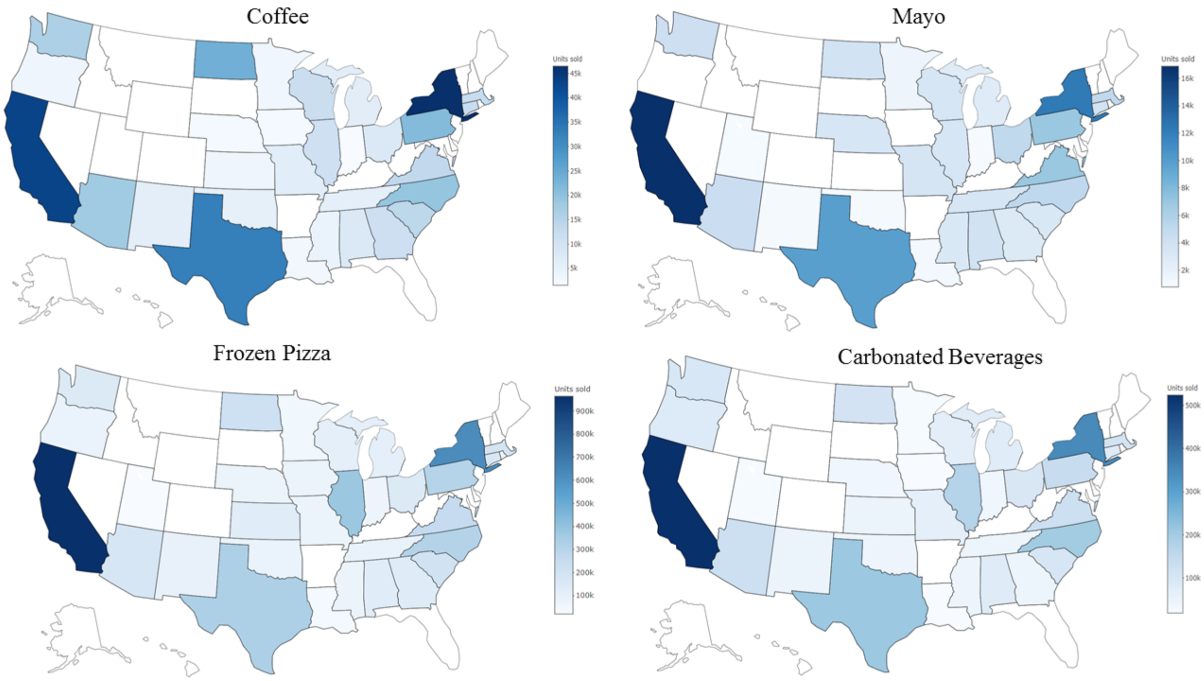
In this section we analyze a part of the dataset published by Bronnenberg et al. (2008). This dataset has been used in significant studies related to consumer behavior, spending and their policy implications (see for example Bronnenberg et al. (2012); Coibion et al. (2015)). The dataset holds the weekly sales and scanner prices of common grocery items sold in retail outlets across 50 states in the U.S. The retail outlets available in the dataset have identifiers that link them to the city that they serve. In accordance to our lagged data example, we analyze a part of this dataset that spans $m = 100$ weeks from December 31, 2007 to November 29, 2009 as substantial amount of disaggregate data from distant past that will be used for constructing auxiliary information on the covariance. We use 3 weeks from a relatively recent snapshot covering October 31, 2011 to November 20, 2011 as data from the current model. We assume, as in equation (6), that there might have been drift change in the sales data across time but the covariances across stores are invariant over time. Our goal is to predict the state level total weekly sales across all retail outlets for four common grocery items: coffee, mayo, frozen pizza and carbonated beverages. We use the most recent $T = 2$ weeks, from November 7, 2011 to November 20, 2011 as our prediction period and utilize the sales data of week $t-1$ to predict the state aggregated totals for week t where $t = 1, \dots, T$. For each of the four products, the prediction period includes sales across approximately $n = 1,140$ retail outlets that vary significantly in terms of their size and quantity sold across the T weeks. Moreover, some of the outlets have undergone

$\mathbb{X}$

0 999	**1 002**
0 995	**1 004**
1 000	**0 998**
1 003	**0 984**

$\sum \quad \mathbb{I}\,\mathbb{X}$

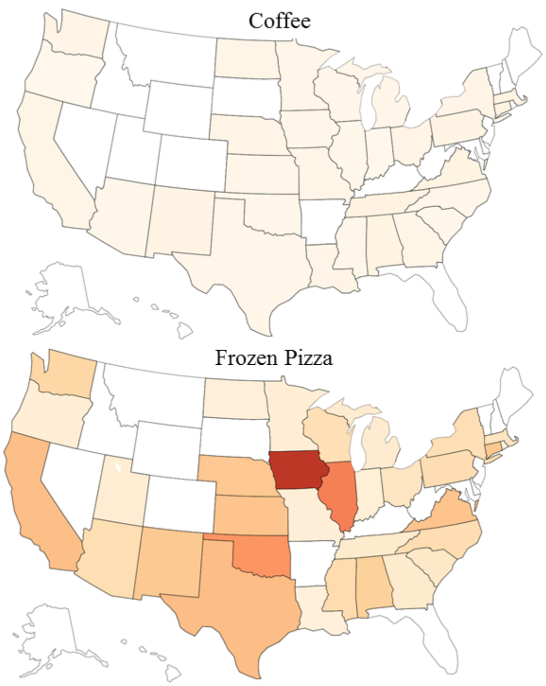
$$\mathcal{L}\left(\qquad \right) \quad \frac{\sum^{p}\left\{ \left(\qquad \right) \quad \left(\qquad \right) \right\}}{\sum^{p}\left\{ \left(\qquad \right) \quad \left(\qquad \right) \right\}}$$



$\mathcal{L}$

`smooth.spline` `splines2`

1


Coffee
Mayo
Frozen Pizza
Carbonated Beverages
1
0.8
0.6
0.4
0.2
0

7. Discussion

low rank plus homoscedastic noise

Appendix A. Proofs

A.1 Preliminary expansions for eigenvector and eigenvalues

$\sqrt{}$

$\mathbb{R}$

$\mathbb{S}$ $\mathbb{R}$

$\mathbb{R}$

$\sqrt{}$

$\sqrt{}$ $\sqrt{}$

$\sqrt{}$

$\mathcal{B}$ $\mathbb{R}$

$\mathcal{B}$

$\sqrt{}$

$\sqrt{}$ $\sqrt{}$

A.2 Proof of Theorem 1A

$\mathbb{R}$

$\mathcal{B}$

$$\sum -$$

$\sum$ $\sqrt{}$

$\sqrt{}$

$\mathcal{B}$

A3

$'$

$\boldsymbol{T}_0$ $\boldsymbol{B}_0$ $\boldsymbol{b}$ $\mathcal{B}$ $\sqrt{}$

A.3 Proof of Lemma 2

$$\sum \qquad \sum$$

A.4 Proof of Theorem 1B

A3

$\mathbb{R}$

$\sum \quad \frac{}{}{}^{2}_{j} \quad \sqrt{}$

$(\quad) \quad (\quad)$

$\sim \left[\sim \quad \sim \right] \sim$

$\sim \quad \sim$

$- \quad - \quad (\quad)$

$(\quad)$

$\sim \quad \sim$

$\sim \quad \sim$

$$\frac{\sqrt{}\sqrt{}}{} \quad \frac{}{}$$

$$\frac{\sqrt{}\sqrt{}}{} \quad \frac{}{}$$

$\sim \quad \sim$

$$\left[\sim \quad \sim \right] \left[\quad \right]$$

$$\left(\quad - \quad \right) \sim \left[\quad - \quad \right]^{\sim}$$

A.5 Proof of Lemma 1

$\mathbb{E}\ {}_i\mathcal{L}$ $\mathcal{L}$

$$\mathbb{E}\ {}_i\mathcal{L} \quad \Big[\quad \Big]$$

$$\mathbb{E}\ {}_i\ \boldsymbol{AX}\ {}_i \quad \Big[\quad \Big]$$

$\mathbb{E}\ {}_i\mathcal{L}$ $\mathbb{E}$

$\mathbb{E}$

A.6 Proof of Lemma 3

A.7 Proofs of Lemmata 4 and 5

Lemma A. *Under assumptions **A1** and **A2**, uniformly in $\mathcal{B}$ such that $\mathcal{B}$ for any fixed , with , and for all $\mathbb{R}$, we have as ,*

$$_{\boldsymbol{b}\ \mathcal{B}} \Big| \quad \mathcal{J} \quad \mathcal{J}$$

$$\sum \quad \Big| \quad \sqrt{}$$

Proof of Lemma A.

$\mathcal{J}$ $\mathcal{J}$

$\mathcal{J}$ $\mathcal{J}$

$\mathcal{B}$ $\mathcal{J}$

$$\sum\sum\sum \frac{\quad '}{\quad '} \quad , \quad ,$$

$$\mathcal{J} \quad \sum\sum \frac{\quad '}{\quad '} \quad , \quad ,$$

$$\sum\sum \frac{}{} \quad \mathcal{J} \quad \sum \frac{}{}$$

$$\sum \quad \mathcal{J}$$

$$\Big(\sum \quad\Big)\Big(\sum \quad\Big)$$

$$\Big(\sum \quad\Big) \quad \mathcal{J} \quad \sum\Big(\quad\Big) \quad \sqrt{\quad}$$

$\mathcal{B}$

Proof of Lemma 4, statement (a) $\mathbb{E}\Big\{\Big(\quad\Big)\Big\}$

$$\mathbb{E}\Big\{\Big(\quad\Big)\Big\} \quad \Big\{\Big(\quad\Big)\Big\}$$

$$\Big[\Big(\quad\Big) \quad \Big(\quad\Big)\Big]$$

$$\sqrt{\quad}$$

$\mathcal{J}$ $\mathcal{J}$ $\mathcal{J}$

$$\frac{\mathcal{J}}{\mathcal{J}}$$

$\mathcal{J}$ $\mathcal{J}$ $\sqrt{}$

$\mathcal{J}$ $\sum$ $\sqrt{}$

$$\frac{\mathcal{J}}{\mathcal{J} \quad \sum} \quad \left(\sqrt{\frac{}{}}\right)$$

$\mathcal{J}$

Proof of Lemma 4, statement (b) $\mathcal{J}$

$\mathcal{J}$ $\sum$ $\sqrt{}$

$\mathcal{J}$ $\sqrt{}$

Proof of Lemma 4, statement (c)

$$\mathbb{E}\Big[\Big(\quad \Big) \Big] \quad \mathbb{E}\Big[\Big(\quad \Big) \Big]$$

Proof of Lemma 5

A.8 Proofs of Theorems 2A and 2B

References

Bulletin of the American Mathematical Society

Journal of Multivariate Analysis

Hierarchical modeling and analysis for spatial data

arXiv preprint arXiv:2008.11903

Journal of Multivariate Analysis

Probability theory and related fields

Marketing science

American Economic Review

The Annals of Statistics

Annals of Statistics

The Annals of Statistics

Cell reports

American Economic Review

Generalized linear models: A Bayesian perspective

Biometrika

Large-scale inference: empirical Bayes methods for estimation, testing, and prediction

Computer age statistical inference

Journal of Econometrics

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

arXiv preprint arXiv:1602.00719

Shrinkage Estimation

CRC, London

Data analysis using regression and multilevel/hierarchical models

Econometric Theory

The Annals of Statistics

Latent variables in socio-economic models

Management Science

Journal of Machine Learning Research

Optimality: The Third Erich L. Lehmann Symposium

The Review of Financial Studies

Journal of the American Statistical Association

Proceedings of the IEEE

arXiv preprint arXiv:1105.1404

Mathematical Programming

Econometrica: journal of the Econometric Society

Journal of the American Statistical Association

Bayesian Analysis

Optimal shrinkage estimation in heteroscedastic hierarchical linear models

Journal of Financial Economics

Chemometrics and Intelligent Laboratory Systems

IEEE Transactions on Signal Processing

The Annals of Statistics

arXiv preprint arXiv:1511.00028

Journal of Econometrics

Statistical Science

Random Matrices: Theory and Applications

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

The Journal of Finance

Journal of Financial Economics

Journal of Econometrics

Statistica Sinica

Journal of Statistical Planning and Inference

Subjective and objective Bayesian statistics: principles, models, and applications

Herbert Robbins Selected Papers

Cell reports

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

Bernoulli

Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage

Journal of the American Statistical Association

Journal of the American Statistical Association

Annals of statistics

Statistica Sinica

Journal of the American Statistical Association

Ann. Statist.