# Improved Shrinkage Prediction under a Spiked Covariance Structure

Trambak Banerjee[*], Gourab Mukherjee[*] and Debashis Paul[†]

September 3, 2018

## Abstract

We develop a novel shrinkage rule for prediction in a high-dimensional non-exchangeable hierarchical Gaussian model with an unknown spiked covariance structure. We propose a family of commutative priors for the mean parameter, governed by a power hyper-parameter, which encompasses from perfect independence to highly dependent scenarios. Corresponding to popular loss functions such as quadratic, generalized absolute, and linex losses, these prior models induce a wide class of shrinkage predictors that involve quadratic forms of smooth functions of the unknown covariance. By using uniformly consistent estimators of these quadratic forms, we propose an efficient procedure for evaluating these predictors which outperforms factor model based direct plug-in approaches. We further improve our predictors by introspecting possible reduction in their variability through a novel coordinate-wise shrinkage policy that only uses covariance level information and can be adaptively tuned using the sample eigen structure. We extend our methodology to aggregation based prescriptive analysis of generic multidimensional linear functionals of the predictors that arise in many contemporary applications involving forecasting decisions on portfolios or combined predictions from dis-aggregative level data. We propose an easy-to-implement functional substitution method for predicting linearly aggregative targets and establish asymptotic optimality of our proposed procedure. We present simulation experiments as well as real data examples illustrating the efficacy of the proposed method.

**Keywords:** Shrinkage predictors, Asymptotic optimality, Spiked Covariance, Non-exchangeable priors, Aggregated predictants, Asymmetric loss, Generalized absolute loss, Linex.

[*] University of Southern California, Los Angeles

[†] University of California, Davis

# 1 Introduction

In every branch of big-data analytics, it is now commonplace to use notions of shrinkage for the construction of robust algorithms and predictors. Over the last decade, driven by applications in a wide range of scientific problems, the traditional roles of statistical shrinkage have rapidly evolved as new perspectives have been introduced to address and exploit complex, latent structural properties of modern datasets. Incorporating such structural properties vastly improves predictive efficiency. Traditional shrinkage estimators in high-dimensional location models (see Brown and Greenshtein (2009), Dicker and Zhao (2016), Efron (2012), Efron and Hastie (2016), Fourdrinier, Strawderman, and Wells (2017), Greenshtein and Park (2009), Greenshtein and Ritov (2009), Koenker and Mizera (2014), Robbins (1985), Zhang (2003) and the references therein) were developed based on homoscedastic models using notions of spherical symmetry. Recent results of Brown, Mukherjee, and Weinstein (2018), Tan (2015), Weinstein, Ma, Brown, and Zhang (2015), Xie, Kou, and Brown (2012, 2016) have brought to light new shrinkage phenomena in heteroscedastic models. However, these results are based on multivariate set-ups with known covariances. For prediction in location models with dependence, it is difficult to assimilate optimal structural properties of covariances using these shrinkage rules. In a host of modern applications in biology, economics, finance, health-care and supply chain management which are briefly narrated below, we need simultaneous predictions of several dependent variables where including domain specific regularization on their covariances is beneficial.

1. In portfolio selection, the vector of next period excess returns on investable assets form a critical component in determining the optimal portfolio weights (Karoui et al., 2011). Prediction programs of different flavors are employed to estimate the future returns with several popular approaches using factor covariance models (Fan, Fan, and Lv, 2008, Johnstone and Titterington, 2009) to capture the dependence among asset returns (Kozak et al., 2017).

2. In cell-biology, the problems of predicting the expressions of several genes leads to inference in a high-dimensional location model (Cavrois et al., 2017, Sen et al., 2014). Effective statistical methods usually integrate the dependence structure of gene expressions while conducting inference on such high-dimensional location parameters (Sun and Cai, 2009).

3. In health-care management, simultaneous prediction of several inventories or resources is very important for optimal operations. For instance, general operations of a health-care provider need simultaneous prediction of the number of nurses that it will be needing in its different hospitals (Green et al., 2013). The loss function for the health-care provider is agglomerative across its different hospitals. Another interesting caveat here is that the loss functions here are asymmetric as

a hospital would incur an underage cost if too many patients arrive and expensive agency nurses have to be called, and an overage cost if too many regular nurses are scheduled compared to the number of patients. In this paper, we study shrinkage prediction under such loss functions. In Mukherjee et al. (2015) it was seen that for such compound decision theoretic problems in uncorrelated models, empirical Bayes induced shrinkage can provide better performance than simple coordinate-wise rules. Incorporating the dependence structure among the patient arrivals in different hospitals would improve shrinkage rules developed for uncorrelated models.

4. A topic of vibrant current research in supply chain management (Levi et al., 2015, Rudin and Vahn, 2014) is the inventory optimization problem of distributors and retailers who, based on past sales data, need to predict future demands and balance the trade-offs between stocking too much and incurring high depreciation costs on unsold inventory versus stocking too little and suffering tremendous reputation and lost sales costs. Here, we study the optimal stocking problem by analyzing grocery sales data across several retail outlets in USA. For any distributor forecasting the future sales across so many outlets translates to a high-dimensional demand prediction problem where incorporating co-dependencies in the demands among different stores is potentially useful.

Often, the data in these highly multivariate applications have approximate low-dimensional representations that can be described through a factor model. Thus, the variability in such data can be well represented through a spiked covariance model (see for example Cai, Ma, and Wu (2013), Dobriban, Leeb, and Singer (2017), El Karoui (2008), Fan, Liao, and Mincheva (2013), Kritchman and Nadler (2009), Ma (2013), Onatski, Moreira, and Hallin (2014) and the references therein). For constructing efficient predictors, it is important to leverage the presence of such covariance structures. However, the well-studied high-dimensionality effects on the eigenvectors and eigenvalues of the sample covariance matrix (Johnstone and Paul, 2018, Johnstone and Titterington, 2009, Onatski, 2015, Paul, 2007) also suggest the need for appropriate regularization even while making use of the spiked covariance structure.

In this article, we propose CASP - a Coordinate-wise Adaptive Shrinkage Prediction rule for shrinkage prediction in high-dimensional Gaussian models with unknown location as well as unknown spiked structured covariances. Motivated by contemporary applications, we consider the set up where we observe only a few observations from the model, but also assume having some auxiliary information on the covariance. We provide a rigorous framework for constructing such auxiliary information based on lagged (see section 2.3) dis-aggregate level data which often arises in research problems based on industrial datasets.

To facilitate a potent and robust notion of shrinkage in such Gaussian models, we consider a hierarchical set-up based on non-exchangeable priors for the mean vector. Our proposed prior structure involves a shape and a scale hyper-parameter, as well as the unknown population covariance, and is designed to describe the structure of the mean vector in terms of its representation in the spectral coordinates of the population covariance. The shape hyper-parameter regulates the contributions of low-variance principal components and produces a wide class of priors ranging from complete independence to highly dependent scenarios. For a range of commonly used loss functions, the Bayes predictors in these hierarchical models involve quadratic forms involving smooth functions of the population covariance. In practice, one needs to estimate these quantities based on available information, which constitutes a core challenge of this formulation.

We work under the framework where the auxiliary information allows us to construct an estimator of the unknown population covariance that has degrees of freedom comparable to the dimensionality of the observations. Then, we make use of the results on the behavior of eigenvalues and eigenvectors of high-dimensional sample covariance matrix (Baik and Silverstein, 2006, Onatski, 2012, Paul, 2007), to develop a bias-correction principle that leads to an efficient approach for evaluating the Bayes predictors. Thereafter, by introducing a novel coordinate-wise shrinkage term we provide additional improvements on the performance of these Bayes predictors. Our proposed CASP methodology systematically assimilates these key features. In addition, we provide a detailed analysis of the operational characteristics of the proposed CASP procedure for both aggregated and dis-aggregated forecasting problems. Our analysis demonstrates that, even for linearly aggregated prediction problems, the substitution-based CASP procedure is still asymptotically optimal as long as the dimension of the aggregation subspace is suitably small compared to dimension of the observations.

The paper is organized as follows. In Section 2, we describe our predictive set-up. In Section 3, our proposed methodology CASP and its asymptotic properties are presented. In section 4, we analyze the performance of CASP in aggregated predictive models. Numerical performances of our methods are investigated using both simulated and real data in Sections 5 and 6 respectively. Proofs and additional technical details are relegated to the Appendix and the supplementary material.

## 2 Predictive setup: Aggregative predictors and Side-information

We first introduce the statistical prediction analysis framework of Aitchison and Dunsmore (1976) and Geisser (1993). Then we give an overview of the high-dimensional orthogonal prediction setup (George et al., 2006, George and Xu, 2008, Mukherjee et al., 2015), and thereafter introduce aggregative prediction objectives based on dis-aggregative orthogonal predictive models.

## 2.1 Predictive Model

Consider an $n$ dimensional Gaussian location model where the observed past $\boldsymbol{X} = (X_1, \dots, X_n)$ as well as the future observation $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ are distributed according to a normal distribution with an unknown mean $\boldsymbol{\theta}$ and unknown covariance proportional to $\boldsymbol{\Sigma}$. The past and the future are related only through the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$, conditioned on which they are independent. The orthogonal predictive model is:

$$\texttt{Past observations } \boldsymbol{X} \sim N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma}), \texttt{ and Future } \boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, m_0^{-1}\boldsymbol{\Sigma}) \,, \tag{1}$$

where, $\boldsymbol{\Sigma}$ is an unknown $n \times n$ positive definite matrix, $\boldsymbol{\theta}$ is an unknown $n \times 1$ vector and $m_0 > 0$ is a known constant (typically $m_0 = 1$). Based on observing $\boldsymbol{x}$, the goal at the the dis-aggregative level is to predict $\boldsymbol{Y}$ by $\hat{\boldsymbol{q}}(\boldsymbol{x})$ under a loss that is agglomerative across the $n$ dimensions.

## 2.2 Linearly Aggregated Predictors

In a host of modern applications, we are interested in predicting several linearly aggregated components from model (1). The predictant here is $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Y}$, where the transformation matrix $\boldsymbol{A} \in \mathbb{R}^{p \times n}$ is observed, with $p \leq n$ and full rank. Instead of prediction at the dis-aggregate level, the goal is to formulate $\hat{\boldsymbol{q}} = \{\hat{q}_i(\boldsymbol{X}) : 1 \leq i \leq p\}$ based on the dis-aggregative past data $\boldsymbol{X}$ such that $\hat{\boldsymbol{q}}$ optimally forecasts $\boldsymbol{V}$. The loss function is cumulative across the $p$ components of $\boldsymbol{V}$.

Aggregative prediction problems of this flavor often arise in several applications. For example, in portfolio selection (Pástor, 2000), $\boldsymbol{A}$ might represent the $p \times n$ portfolio weight matrix of $p$ investors and $\boldsymbol{Y}$ is the next period excess return vector on the $n$ assets. Similarly, as discussed in section 6, in supply chain management distributors may need to forecast the future sales of their products across a large number of retail outlets spread over various locations or states. Often the high inter-state transfer costs forbid the distributors to deliver their products to these retail outlets from a central warehouse. Instead, the products are typically sourced at regional or state warehouses which are then distributed to the retail outlets in the nearby region. In this demand forecasting setup then, the matrix $\boldsymbol{A}$ might represent the $p \times n$ aggregation matrix that aggregates the demand for each product across the $n$ retail outlets into $p$ states and $\boldsymbol{Y}$ is the $n$ dimensional future demand vector at each retail outlet. Such problems, where the target distribution is different from that of past observations, are more challenging than dis-aggregate level prediction (George and Xu, 2008, Komaki, 2015, Yano and Komaki, Yano and Komaki). Naturally, in this set-up when $p = n$ and $\boldsymbol{A} = \boldsymbol{I}_n$, we revert to prediction at dis-aggregate level.

Model (1) is a natural extension of the uncorrelated heteroscedastic model of Xie, Kou, and Brown

(2012). For correlated and known $\boldsymbol{\Sigma}$, shrinkage estimation in (1) for quadratic loss has been studied in Kong et al. (2017). However, the parameters of interest studied in multi-level models are correlated in most practical situations. The correlation structure is usually unknown and requires estimation. Here, we consider shrinkage prediction in (1) when $\boldsymbol{\Sigma}$ is estimated from observations $\boldsymbol{W}_j = (W_{1j}, \ldots, W_{nj})^T$ where $\boldsymbol{W}_j|\boldsymbol{\mu}_j$ are independently distributed from $N_n(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ for $j = 1, \ldots, m$. Given the parameters $\boldsymbol{\theta}, \boldsymbol{\Sigma}$ and $\boldsymbol{\mu} := \{\boldsymbol{\mu}_j : 1 \leq j \leq m\}$ in our predictive set-up, $\boldsymbol{X}$, $\boldsymbol{W} := [\boldsymbol{W}_1, \ldots, \boldsymbol{W}_m]$, and $\boldsymbol{Y}$ are independent.

In many real world applications $\boldsymbol{\mu_j}$'s are very different from $\boldsymbol{\theta}$, in which cases $\boldsymbol{W}$ provides information about $\boldsymbol{\Sigma}$ but not much on $\boldsymbol{\theta}$. For instance, there may have been a drift in the data generation process over time without affecting the correlation structures. In such rapid trend changing environments (Harvey et al., 2016, Kozak et al., 2017, Patton and Timmermann, 2007), there often exist related instruments that can be used to estimate the covariances but not the average. In the asset pricing context for estimating the joint explanatory power of a large number of cross-sectional stock return predictors, Kozak et al. (2017) suggest using the daily returns to estimate covariances while conducting shrinkage to regulate the uncertainty about means. Most datasets available for research from the industry contain dis-aggregate data from lagged past as data from immediate past could potentially reveal their current operational strategies. Prediction in such problems involve high-dimensional correlated location model (1) where one or very few (in which case they can be summarized into $\boldsymbol{X}$ by taking average) immediate past observation vectors are available. In the following section, we show that for prediction based on such high-dimensional lagged datasets, we can construct side information $\boldsymbol{W}$.

### 2.3 Construction of suitable side-information regarding Covariance

Consider observing $\boldsymbol{W}_t$ for $t = t_0 + 1, \ldots, t_0 + m$ time periods from the drift changing model

$$\boldsymbol{W}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t \tag{2}$$

where $\boldsymbol{\epsilon}_t \overset{i.id}{\sim} N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$. Let $\boldsymbol{X}$ be a vector from this model from the most recent time period so that $\boldsymbol{X} = \boldsymbol{\mu}_{t_c} + \boldsymbol{\epsilon}_{t_c}$ with $\boldsymbol{\mu}_{t_c}$ much different than $\{\boldsymbol{\mu}_t : t = t_0 + 1, \ldots, t_0 + m\}$ as the time lag $t_c - t_0 - m$ is huge. We are interested in predicting the future vector $\boldsymbol{Y}$ from the next time period $t_c + 1$ where $\boldsymbol{Y} = \boldsymbol{\mu}_{t_c+1} + \boldsymbol{\epsilon}_{t_c+1}$. Compared to the variability produced by $\boldsymbol{\Sigma}$, the difference $||\boldsymbol{\mu}_{t_c} - \boldsymbol{\mu}_{t_c+1}||$ can be ignored. Let $\boldsymbol{\theta} = \boldsymbol{\mu}_{t_c}$ and as $\boldsymbol{\epsilon}_t \overset{i.i.d}{\sim} N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ for $t \in \{t_0+1, \ldots, t_0+m\}$, it can be considered that both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are from model (1). Note, as $\boldsymbol{X}$ has only one past vector and $m$ is large , it does not benefit us much to use $\boldsymbol{X}$ for estimating $\boldsymbol{\Sigma}$. Throughout this paper, we will estimate functionals involving $\boldsymbol{\Sigma}$

based only on $\boldsymbol{W}$, whereas our predictors will involve $\boldsymbol{X}$ as it contains pivotal information about the current drift $\boldsymbol{\theta}$. Given the parameters $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m, \boldsymbol{\theta}, \boldsymbol{\Sigma})$, $\boldsymbol{X}$, $\boldsymbol{W} := [\boldsymbol{W}_{t_0+1}, \dots, \boldsymbol{W}_{t_0+m}]$, and $\boldsymbol{Y}$ are independent with the goal being to predict $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{Y}$ by using the information in $\boldsymbol{W}$ and $\boldsymbol{X}$.

It is observed in Banerjee et al. (2017), Binkiewicz et al. (2014), Cai et al. (2016), Ke et al. (2014) that methods which successfully leverage associated auxiliary information vastly enhance estimation accuracy. We next describe an efficient methodology to extract suitable information about $\boldsymbol{\Sigma}$ from $\boldsymbol{W}$, which we will treat as side information on $\boldsymbol{\Sigma}$ and treat $\boldsymbol{X}$ as primary information on the current model. Often $\boldsymbol{\mu}_t$ can be well approximated by low dimensional processes. We use $k$ basis functions to model $\boldsymbol{\mu}_t$ over time and let

$$\boldsymbol{W}_t = \boldsymbol{U}\boldsymbol{C}_t + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}_t \text{ where, } t = t_0+1, \dots, t_0+m,$$

and $\boldsymbol{C}_t$ is a $k \times 1$ vector of basis coefficients, $\boldsymbol{U} \in \mathbb{R}^{n \times k}$ is the matrix of unknown coefficients and $\boldsymbol{\delta}_t \overset{i.i.d}{\sim} N_n(\boldsymbol{0}, \boldsymbol{I}_n)$. In matrix notations, the above can be expressed as $\boldsymbol{W}^T = \boldsymbol{\mathcal{C}}\boldsymbol{U}^T + \boldsymbol{\Delta}\boldsymbol{\Sigma}^{1/2}$ where $\boldsymbol{\mathcal{C}}_{m\times k} = (\boldsymbol{C}_{t_0+1}, \cdots, \boldsymbol{C}_{t_0+m})^T$ and $\boldsymbol{\Delta}_{m\times n} = (\boldsymbol{\delta}_{t_0+1}, \cdots, \boldsymbol{\delta}_{t_0+m})^T$. Consider the projection matrix $\boldsymbol{P}_c = \boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{C}}^T\boldsymbol{\mathcal{C}})^{-1}\boldsymbol{\mathcal{C}}^T$. Then , note that $\boldsymbol{S} = \boldsymbol{W}(\boldsymbol{I}_m - \boldsymbol{P}_c)\boldsymbol{W}^T$ follows an $n$-dimensional Wishart distribution with degrees of freedom $m_k = m - k$ and covariance $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}_m - \boldsymbol{P}_c)\boldsymbol{\Sigma}^{1/2}$. Here $m$ is large and $k$ is much smaller than $m$, and so we expect $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{I}_m - \boldsymbol{P}_c)\boldsymbol{\Sigma}^{1/2}$ to be a good approximation to $\boldsymbol{\Sigma}$. Henceforth, without loss of generality but with a slight abuse of notation which literally replaces $m_k$ by $m$, we assume that $\boldsymbol{S}$ follows $\mathsf{Wishart}_n(m, \boldsymbol{\Sigma})$.

## 2.4 Spiked Covariance structure

We further assume that the unknown covariance has a spike covariance structure (Baik and Silverstein, 2006, Johnstone and Lu, 2012, Paul and Aue, 2014). It is popularly used to model signal-plus-noise decomposition of the centered (mean-subtracted) observations, where the signal belongs to a low-dimensional linear subspace and noise is isotropic (Benaych-Georges and Nadakuditi, 2012, Passemier and Yao, 2012). Suppose $\boldsymbol{\Sigma}$ in model (1) is of the form:

$$\boldsymbol{\Sigma} = \sum_{j=1}^{K} \ell_j \mathbf{p}_j \mathbf{p}_j^T + \ell_0 (I - \sum_{j=1}^{K} \mathbf{p}_j \mathbf{p}_j^T) \tag{3}$$

where, $\mathbf{p}_1, \dots, \mathbf{p}_k$ are orthonormal, $\ell_1 > \cdots > \ell_K > \ell_0 > 0$ and the number of spikes $1 \le K \ll n$. Let the spectral decomposition of $\boldsymbol{S}$ be $\sum_{j=1}^{K} \hat{\ell}_j \hat{\mathbf{p}}_j \hat{\mathbf{p}}_j^T$ where $\hat{\mathbf{p}}_j$ are orthonormal and $\hat{\ell}_1 \ge \cdots \ge \hat{\ell}_n$.

A key result in this model is the so-called *phase transition of the eigenvalues* of $\mathbf{S}$, when $n/m \to \rho > 0$ as $n \to \infty$ and $K$ is fixed (Baik and Silverstein, 2006, Benaych-Georges and Nadakuditi, 2012, Onatski, 2012, Paul, 2007). There is also a corresponding phase transition result for the corresponding eigenvectors of $\mathbf{S}$ (Onatski, 2012, Paul, 2007). Based on these results, for significant spikes, the principal eigenvectors and eigenvalues can be consistently estimated if $K$ is well estimated. This has been the approach taken by Kritchman and Nadler (2008, 2009), Passemier and Yao (2012), and more recently, by Passemier et al. (2015). Our proposed predictive rule uses a similar strategy to deal with spike structures and conducts uniform estimation of quadratic forms involving smooth functions of $\boldsymbol{\Sigma}$ by appropriately adjusting the sample eigenvalues and eigenvectors (see Section 3.1). Specifically, we aim to make use of known results on the behavior of sample eigenvalues and eigenvectors, to develop a simple *substitution principle* that leads to consistent estimators of linear functionals of population eigenvectors and quadratic function of smooth functions of the population covariance matrix.

## 2.5 Hierarchical Modeling with Non-exchangeable priors

We consider prediction in correlated hierarchical models, which result in non-exchangeability of the corresponding coordinate problems. Hierarchical modeling provides an effective tool for combining information and achieving partial pooling of inference (Kou and Yang, 2015, Xie et al., 2012). Though the hierarchical framework allows usage of exchangeable as well as nonexchangeable priors, traditionally shrinkage algorithms in this framework have been developed under exchangeable priors (Fourdrinier et al., 2017, Zhang, 2003). However, in many contemporary applications involving correlated Gaussian models, we need non-exchangeable priors to suitably incorporate auxiliary information regarding covariances in the hierarchical framework (Harvey et al., 2016, Pástor, 2000, Pástor and Stambaugh, 2000). Here, we impose a class of commutative conjugate priors with power-decay on the location parameter $\boldsymbol{\theta}$, which is related to the unknown covariance by hyper-parameters $\beta$, $\tau$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Sigma}, \boldsymbol{\eta}, \tau, \beta) \sim N_n\Big(\boldsymbol{\eta},\ \tau\,\boldsymbol{\Sigma}^{\beta}\Big) \tag{4}$$

The shape parameter $\beta$ is key to controlling the essential characteristic of the posterior density of $\boldsymbol{\theta}$ under model (1). As $\beta$ varies in $[0, \infty)$, it produces a large family of priors capable of reflecting perfect independence to highly dependent scenarios. When $\beta = 0$, the exchangeable prior on the locations resembles the set-up of Xie et al. (2012) with known diagonal covariance. With $\beta = 1$ the prior has the same correlation structure as the data whereas with $\beta > 1$ the prior is relatively less concentrated towards the dominant variability directions in the data. In the finance literature, this family of priors is widely used in asset pricing for formulating varied economically motivated priors that induce shrinkage estimation of market factors (Kozak et al., 2017). While $\beta = 0$ corresponds to the diffuse prior in Harvey et al. (2016), $\beta = 1$ gives the asset pricing prior in Pástor (2000), Pástor and Stambaugh (2000) and $\beta = 2$ yields the prior proposed in Kozak et al. (2017) that shrinks the contributions of low-variance principal components of the candidate factors. The $\mathsf{scale}$ parameter $\tau$ is allowed to vary between 0 to $\infty$. The location parameter $\boldsymbol{\eta}$ is usually restricted to some pre-specified low dimensional subspace. For instance, we can restrict $\boldsymbol{\eta}$ to a 1-dimensional sub-space and consider $\boldsymbol{\eta} = \eta\mathbf{1}$ and estimate $\eta$ as a hyper-parameter. For simplicity, we consider $\boldsymbol{\eta}$ to be set to a pre-specified value $\boldsymbol{\eta}_0$. For the lagged data model of equation (2), we can set $\boldsymbol{\eta}$ to $\boldsymbol{W}\boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{C}}^T\boldsymbol{\mathcal{C}})^{-1}C_{\mathsf{new}}$ or to the grand mean across coordinates $n^{-1}\mathbf{1}^T\boldsymbol{W}\boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{C}}^T\boldsymbol{\mathcal{C}})^{-1}C_{\mathsf{new}}$ where $C_{\mathsf{new}}$ is a known vector of basis coefficients. Our goal is to construct predictive rules for aggregated predictants $\boldsymbol{V}$ under different popular loss functions in the hierarchical model governed by (1) and (4).

## 2.6 Popular loss functions and Bayes Predictors

In this article, we consider popular loss functions that routinely arise in applications - quadratic loss, $\ell_1$ loss, generalized absolute loss and Linex loss. While the quadratic loss is the most widely studied loss function in Statistics, the utility and necessity of asymmetric losses, like the generalized absolute loss and Linex loss, has long been acknowledged, for instance in the early works of Granger (1969), Koenker and Bassett Jr (1978), Zellner (1986), Zellner and Geisel (1968). In what follows, we briefly discuss the two aforementioned asymmetric losses and then present the Bayes predictors for the loss functions considered in this paper.

*Generalized absolute loss function* - also referred to as the check loss (see Chapter 11.2.3 of Press (2009)), is a piecewise linear loss function with two linear segments and uses differential linear weights to measure the amount of underestimation or overestimation. It is the simplest as well as the most popular asymmetric loss function and is fundamental in quantile regression (Koenker and Bassett Jr, 1978). If $\hat{q}_i(\boldsymbol{X})$ represents the predictive estimate of the future $V_i$, then under generalized absolute loss the $i^{th}$ coordinate incurs a loss

$$\mathcal{L}_i(V_i, \hat{q}_i(\boldsymbol{x})) = b_i(V_i - \hat{q}_i)^+ + h_i(\hat{q}_i - V_i)^+ \tag{5}$$

where $b_i$ and $h_i$ are known positive costs associated with underestimation and overestimation respectively in coordinate $i$. In inventory management problems (Levi et al., 2015, Mukherjee et al., 2015, Rudin and Vahn, 2014) for example, where overestimation leads to depreciation and storage costs, but underestimation may lead to significant reputation costs for the retailers, the generalized absolute loss function arises naturally with $b_i \gg h_i$.

*Linex loss function* - of Varian (1975), on the other hand, uses a combination of linear and exponential functions (and hence its name) to measure errors in the two different directions. The loss associated with coordinate $i$ is

$$\mathcal{L}_i(V_i, \hat{q}_i(\boldsymbol{x})) = b_i\Big\{e^{a_i(\hat{q}_i - V_i)} - a_i(\hat{q}_i - V_i) - 1\Big\} \tag{6}$$

where $a_i \neq 0, b_i > 0$ for all $i$. This loss function is more appropriate for event analysis such as predicting accident counts or crime rates, underestimations of which result in much graver consequences than overestimations, however for small values of $|a|$, linex loss behaves approximately like a quadratic loss function (Zellner, 1986).

To facilitate the ease of presentation, we define a few notations first which will be used throughout the paper. Let $l_p(\boldsymbol{V}, \hat{\boldsymbol{q}}) = p^{-1}\sum_{i=1}^{p}\mathcal{L}_i(V_i, \hat{q}_i)$ denote the average loss for predicting $\boldsymbol{V}$ using $\hat{\boldsymbol{q}}$ which only depends on $\boldsymbol{X}$ and $\boldsymbol{S}$. For each $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{S} = \boldsymbol{s}$, the associated *predictive loss* is $L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}}) = \mathbb{E}_{\boldsymbol{V}} l_p(\boldsymbol{V}, \hat{\boldsymbol{q}})$ where $\boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{\theta}$ and the expectation is taken over the distribution of the future $\boldsymbol{V}$ only. The *predictive risk* is given by $\mathbb{E}_{\boldsymbol{X},\boldsymbol{S}} L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ which, by sufficiency, reduces to $R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}}) = \mathbb{E}_{\boldsymbol{A}\boldsymbol{X}} L_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ wherein the expectation is taken over the distribution of $\boldsymbol{A}\boldsymbol{X}$. Note that the expectation over $\boldsymbol{V}$ is already included in $L_p$. Let $\tilde{b}_i = b_i/(b_i + h_i)$ in the generalized absolute loss function, $\check{\boldsymbol{\Sigma}}_\beta = \boldsymbol{A}\boldsymbol{\Sigma}^\beta\boldsymbol{A}^T$ and define

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

where the dependence of $G_{r,\alpha,\beta}$ on $\tau$ has been kept implicit for notational ease. In particular, when $\boldsymbol{A} = \boldsymbol{I}_n$ then $G_{r,\alpha,\beta} = H_{r,\alpha,\beta}$ where

$$H_{r,\alpha,\beta} \coloneqq H_{r,\alpha,\beta}(\boldsymbol{\Sigma}) = (\boldsymbol{\Sigma}^{-1} + \tau^{-1}\boldsymbol{\Sigma}^{-\beta})^{-r}\boldsymbol{\Sigma}^\alpha.$$

Moreover, when $\boldsymbol{A}$ is a general $p \times n$ rectangular matrix then one may express $G_{r,\alpha,\beta}$ in terms of $H_{r,\alpha,\beta}$ as follows

$$\tau^{-r}G_{r,\alpha,\beta} = \Big\{\boldsymbol{A}H_{0,\beta,0}\boldsymbol{A}^T\Big[\boldsymbol{A}\Big(\tau H_{0,\beta,0} + H_{0,1,0}\Big)\boldsymbol{A}^T\Big]^{-1}\boldsymbol{A}H_{0,1,0}\boldsymbol{A}^T\Big\}^r\Big(\boldsymbol{A}H_{0,1,0}\boldsymbol{A}^T\Big)^\alpha \tag{7}$$

Our goal is to minimize $R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})$ over the class of estimators $\hat{\boldsymbol{q}}$ for all values of $\boldsymbol{\psi}$. An essential intermediate quantity in that direction is the Bayes predictive rule $\boldsymbol{q}^{\mathsf{Bayes}}$ which is the unique minimizer of the integrated Bayes risk $B_p(\tau, \beta) = \int R_p(\boldsymbol{\psi}, \hat{\boldsymbol{q}})\pi(\boldsymbol{\psi}|\boldsymbol{\Sigma}, \tau, \beta)d\boldsymbol{\psi}$ and Lemma 1 below provides the univariate Bayes estimator $q_i^{\mathsf{Bayes}}$ for the loss functions discussed earlier.

**Lemma 1** (Univariate Bayes Estimator)**.** *Consider the hierarchical model in equations* (1) *and* (4)*. If* $\boldsymbol{\Sigma}$

*were known, the unique minimizer of the integrated Bayes risk for coordinate $i$ is*

$$\boldsymbol{q}_i^{\mathsf{Bayes}}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{\Sigma},\tau,\beta) = \mathbf{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0 + \mathbf{e}_i^T G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\eta}_0) + \mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma},\boldsymbol{A},\tau,\beta)$$

*for $i = 1,\ldots,p$, where*

$$\mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma},\boldsymbol{A},\tau,\beta) = \begin{cases} \Phi^{-1}(\tilde{b}_i)\Big(\mathbf{e}_i^T G_{1,0,\beta}\mathbf{e}_i + m_0^{-1}\mathbf{e}_i^T G_{0,1,0}\mathbf{e}_i\Big)^{1/2}, & \textit{for generalized absolute loss} \\ -\dfrac{a_i}{2}\Big(\mathbf{e}_i^T G_{1,0,\beta}\mathbf{e}_i + m_0^{-1}\mathbf{e}_i^T G_{0,1,0}\mathbf{e}_i\Big), & \textit{for linex loss} \\ 0, & \textit{for quadratic loss}. \end{cases}$$

Since $\boldsymbol{\Sigma}$ is unknown, perhaps the simplest approach will be to plug in the sample covariance matrix $\boldsymbol{S}$ in $G_{r,\alpha,\beta}$ that appears in the expression of the Bayes estimator. However, this produces a biased, sub-optimal predictor even when $m$ is comparable in magnitude to $n$. In section 3 we describe an efficient methodology for evaluating the Bayes predictive rules.

## 2.7 Discussion

We develop a new algorithm for prediction in high dimensional Gaussian models with unknown spiked covariances. Motivated by modern research problems involving dis-aggregate level lagged data in drift changing applications (described in section 2.3), our prediction framework involves observing only a few observation vectors from the current model which are summarized in a past vector. We construct useful auxiliary or side information on the unknown covariance from the lagged data. Based on these, we propose a flexible shrinkage methodology CASP that integrates available information on the unknown covariance optimally in determining shrinkage directions. The proposed methodology makes several contributions. To summarize, the key features of CASP are:

- It can be used for prediction under popular symmetric as well as asymmetric losses such as linex and generalized absolute loss. Asymmetric losses arise in modern health care and supply chain management problems and shrinkage prediction under them differs in fundamental aspects from shrinkage under symmetric losses. CASP improves upon recent shrinkage algorithms developed in Mukherjee et al. (2015) for asymmetric losses by optimally integrating covariance information.

- It utilizes the phase transition phenomenon of the sample eigenvalues and eigenvectors seen in spiked covariance models (Onatski, 2012, Paul, 2007) and improves upon naive factor model based methodology by using bias corrected efficient estimates of quadratic forms involving unknown covariance matrix (see Sections 3.1 and 5).

- It is developed based on an hierarchical framework that encompass a wide family of non-exchangeable priors that are used in real world applications (Harvey et al., 2016, Kozak et al., 2017). The prior family involves a shape hyper-parameter that regulates the contributions of low-variance principal components. It makes CASP a very flexible shrinkage method which includes the exchangeable priors set-ups of Xie, Kou, and Brown (2012) as a special case.

- It uses a novel coordinate-wise shrinkage policy that only relies on the covariance level information and introduces possible reduction in the variability of the proposed rules (see section 3.2). Robust data driven schemes are employed to adaptively tune the hyper-parameters. We provide a detailed asymptotic analysis on the scope of improvement due to this coordinate-wise shrinkage policy and establish asymptotic optimality of our proposed procedure (see Lemma 3 and Theorem 2A).

- It can be used for prediction in aggregated models. In a lot of applications, we need to forecast linearly combined measurements by dis-aggregate level data. Aggregative set-ups present several challenges. Unlike dis-aggregative level prediction, here the family of prior in (4) does not commute with the unknown covariance. CASP is built on a simple substitution principle that we establish is asymptotically efficient in such aggregative set-ups (see section 4).

- Across varied simulation regimes, we witness that the improvement in CASP due to incorporation of bias-correction and coordinate-wise shrinkage policies is not just technical but is essential as CASP vastly out-performs other competing shrinkage methods (see section 5).

# 3 Proposed Methodology and Asymptotic Properties

In this section we describe our proposed methodology for the efficient evaluation of the Bayes predictive rules in Lemma 1 (Section 3.1) along with the asymptotic properties of the proposed methodology and thereafter discuss the potential improvement in predictive efficiency brought about by coordinate-wise shrinkage (Section 3.2).

## 3.1 Evaluating Bayes predictors in dis-aggregative models

Since $\boldsymbol{\Sigma}$ is unknown, we need to evaluate the Bayes predictive rules based on $\boldsymbol{X}$ and $\boldsymbol{S}$ only which essentially reduces to estimating the quadratic forms $\boldsymbol{b}^T G_{r,\alpha,\beta}\boldsymbol{b}$ uniformly well for all $\tau, \beta$ where $\boldsymbol{b}$ are known vectors on the $n$ dimensional unit sphere $\mathbb{S}_{n-1}$. In particular, for the dis-aggregative model

($\boldsymbol{A} = \boldsymbol{I}_n$), estimating these quadratic forms involving $G_{r,\alpha,\beta}$ reduces to estimating quadratic forms involving $H_{r,\alpha,\beta}$ which is relatively easier. So, we describe our procedure first for the simpler case of the dis-aggregative model and thereafter present the case of the aggregative model in section 4.

We assume the following asymptotic conditions throughout the paper:

**A1 Asymptotic regime :** Suppose that $\rho_n = \frac{n}{m-1} \to \rho \in (0, \infty)$ as $n \to \infty$.

**A2 Significant spike:** Suppose that $\ell_K > \ell_0(1 + \sqrt{\rho})$.

We next present efficient estimates $\{\hat{\ell}^{\mathsf{e}}_j\}_{j=0}^K$ of the dominating eigenvalues $\{\ell_j\}_{j=0}^K$ of $\boldsymbol{\Sigma}$. Define

$$\zeta(x, \rho) = \left[\frac{1 - \rho/(x-1)^2}{1 + \rho/(x-1)}\right]^{1/2} \text{ with } \zeta_j = \zeta(\ell_j/\ell_0, \rho)$$

Recall that under assumptions **A1** and **A2** the leading eigenvectors and eigenvalues of $\boldsymbol{S}$ have the following properties (Paul, 2007)

$$\hat{\ell}_j - \ell_j \left(1 + \frac{\rho}{(\ell_j/\ell_0 - 1)}\right) = O_P(n^{-1/2}) \ , \ \ j = 1, \dots, K, \tag{8}$$

and

$$\hat{\mathbf{p}}_j \approx \zeta_j \mathbf{p}_j + \sqrt{1 - \zeta_j^2}(I - \mathbf{P}_K \mathbf{P}_K^T) \frac{\boldsymbol{\varepsilon}_j}{\sqrt{n-K}} \ , \ \ j = 1, \dots, K, \tag{9}$$

where $\mathbf{P}_K = [\mathbf{p}_1 : \cdots : \mathbf{p}_K]$ and $\boldsymbol{\varepsilon}_j \sim N(0, I_{n-K})$. We will use these properties to ensure that the quadratic forms of the type $\boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b}$ are consistently estimated. When $K$, the number of significant spikes, is known, we have efficient estimates $\hat{\ell}^{\mathsf{e}}_j$ of $\ell_j$ for $j = 0, \dots, K$ (Passemier et al., 2017) that involve bias correction of $\hat{\ell}_j$ using the approximation properties of equation (8) as follows: Let, $\hat{\ell}_0 = (n-K)^{-1} \sum_{j=K+1}^n \hat{\ell}_j$ and then for $j = 1, \dots, K$, let $\hat{\ell}'_j$ be the solution of the following equation (for $x$)

$$\hat{\ell}_j = \hat{\ell}_0 \psi(x/\hat{\ell}_0, \rho_n) = x\left(1 + \frac{\rho_n}{x/\hat{\ell}_0 - 1}\right).$$

Then, the estimates of $\{\ell_j\}_{j=0}^K$ are $\{\hat{\ell}^{\mathsf{e}}_j\}_{j=0}^K$ where

$$\hat{\ell}^{\mathsf{e}}_0 = \hat{\ell}_0 \left(1 + \frac{\rho_n \hat{\xi}_0}{n-K}\right) \tag{10}$$

and for $j = 1, \dots, K$,

$$\hat{\ell}^{\mathsf{e}}_j = \frac{\hat{\ell}^{\mathsf{e}}_0}{2}\left[(\hat{\ell}_j/\hat{\ell}^{\mathsf{e}}_0 + 1 - \rho_n) + \left((\hat{\ell}_j/\hat{\ell}^{\mathsf{e}}_0 + 1 - \rho_n)^2 - 4\hat{\ell}_j/\hat{\ell}^{\mathsf{e}}_0\right)^{1/2}\right], \tag{11}$$

with $\hat{\xi}_0 = K + \sum_{j=1}^{K} \left(\hat{\ell}'_j / \hat{\ell}_0 - 1\right)^{-1}$. Now, consider the following as an estimate for $H_{r,\alpha,\beta}$

$$\begin{aligned} \hat{H}_{r,\alpha,\beta} &= \sum_{j=1}^{K} \frac{1}{\hat{\zeta}_j^2}(h_{r,\alpha,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}}))\hat{\mathbf{p}}_j\hat{\mathbf{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})I \\ &= \sum_{j=1}^{K}\left[\frac{1}{\hat{\zeta}_j^2}h_{r,\alpha,\beta}(\hat{\ell}_j^{\mathsf{e}}) + \left(1-\frac{1}{\hat{\zeta}_j^2}\right)h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})\right]\hat{\mathbf{p}}_j\hat{\mathbf{p}}_j^T + h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathsf{e}})\left(I - \sum_{j=1}^{K}\hat{\mathbf{p}}_j\hat{\mathbf{p}}_j^T\right) \end{aligned} \tag{12}$$

where $h_{r,\alpha,\beta}(x) = (x^{-1} + \tau^{-1}x^{-\beta})^{-r}x^{\alpha}$ is the scalar version of $H_{r,\alpha.\beta}$, $\hat{\zeta}_j = \zeta(\hat{\ell}_j^{\mathsf{e}}/\hat{\ell}_0^{\mathsf{e}}, \rho_n)$, $\hat{\ell}_0^{\mathsf{e}}$, $\hat{\ell}_j^{\mathsf{e}}$ are from equations (10), (11) respectively and $\hat{\mathbf{p}}_j$ are from equation (9). A key aspect regarding the estimates $\hat{H}_{r,\alpha,\beta}$ in equation (12) is that they not only involve asymptotic adjustments to the sample eigenvalues through equations (10) and (11) but also use the phase transition phenomenon of the sample eigenvectors to appropriately adjust them through $\hat{\zeta}_j$ and equation (9).

The following condition ensures that the results on the behavior of the Bayes predictors and their estimated versions remain valid uniformly over a collection of hyper-parameters.

**A3** $\tau \in \mathbf{T}_0$ and $\beta \in \mathbf{B}_0$ where $\mathbf{T}_0$ and $\mathbf{B}_0$ are compact subsets of $(0,\infty)$ and $[0,\infty)$, respectively.

Notice that **A3** implies in particular that $\tau_0 \le \tau < \infty$ for some $\tau_0 > 0$. For the dis-aggregative model, Theorem 1A proves the asymptotic consistency of $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b}$ uniformly over the hyper-parameters $(\tau, \beta)$ and known vectors $\boldsymbol{b}$ on the $n$ dimensional unit sphere $\mathbb{S}_{n-1}$.

**Theorem 1A** (Asymptotic consistency of $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b}$)**.** *Under assumptions **A1**, **A2**, and **A3**, uniformly over $\tau \in \mathbf{T}_0$, $\beta \in \mathbf{B}_0$ 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 $(r,\alpha) \in \{-1,0,1\} \times \mathbb{R}$,*

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

*where the dependence of $H_{r,\alpha,\beta}$ on $\tau$ has been kept implicit for notational ease.*

An important consequence of Theorem 1A is that it allows us, almost immediately, to construct an efficient evaluation scheme for the Bayes predictive rule in Lemma 1 under the dis-aggregative model as follows:

**Definition 1** (Predictive rule - dis-aggregative model)**.** *Under the hierarchical model of equations (1) and (4), the proposed predictive rule for the dis-aggregative model is given by $\hat{q}^{\mathsf{approx}}$ which is defined as*

$$\hat{q}_i^{\mathsf{approx}}(\boldsymbol{X}|\boldsymbol{S},\tau,\beta) = \mathbf{e}_i^T\boldsymbol{\eta}_0 + \mathbf{e}_i^T\hat{H}_{1,-1,\beta}(\boldsymbol{X}-\boldsymbol{\eta}_0) + \hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S},\tau,\beta) \tag{13}$$

*where*

$$\hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S},\tau,\beta)=\begin{cases}\Phi^{-1}(\tilde{b}_i)\Big(\mathbf{e}_i^T\hat{H}_{1,0,\beta}\mathbf{e}_i+m_0^{-1}\mathbf{e}_i^T\hat{H}_{0,1,0}\mathbf{e}_i\Big)^{1/2}, & \textit{for generalized absolute loss}\\ -\dfrac{a_i}{2}\Big(\mathbf{e}_i^T\hat{H}_{1,0,\beta}\mathbf{e}_i+m_0^{-1}\mathbf{e}_i^T\hat{H}_{0,1,0}\mathbf{e}_i\Big), & \textit{for linex loss}\\ 0, & \textit{for quadratic loss.}\end{cases}$$

Note that $\hat{q}^{\mathsf{approx}}$ is a simple approximation to the Bayes predictive rules in Lemma 1 where, under the dis-aggregative model, the quadratic forms $\boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{b}$ are replaced by their consistent estimates $\boldsymbol{b}^T\hat{H}_{r,\alpha,\beta}\boldsymbol{b}$ in equation (12). Except the second term in (13), all the other estimated quadratic forms are symmetric. The asymmetric quadratic form$\boldsymbol{b}^T H_{r,\alpha,\beta}\boldsymbol{c}$ (where $\boldsymbol{b}$, $\boldsymbol{c}$ are unit vectors) can also be written as a difference of two symmetric quadratic forms $(1/4)\{(\boldsymbol{b}+\boldsymbol{c})^T H_{r,\alpha,\beta}(\boldsymbol{b}+\boldsymbol{c})-(\boldsymbol{b}-\boldsymbol{c})^T H_{r,\alpha,\beta}(\boldsymbol{b}-\boldsymbol{c})\}$, and Theorem 1A can be directly applied to yield Lemma 2 which provides decision theoretic guarantees on the predictors. It shows that uniformly over $(\tau,\beta)$ the largest coordinate-wise gap between $\hat{q}^{\mathsf{approx}}$ and $\boldsymbol{q}^{\mathsf{Bayes}}$ is asymptotically small.

**Lemma 2.** *Under assumptions **A1**, **A2** and **A3**, uniformly over $\tau\in\mathbf{T}_0$, $\beta\in\mathbf{B}_0$, for all $(r,\alpha)\in\{-1,0,1\}\times\mathbb{R}$, we have, conditionally on $\boldsymbol{X}$,*

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

While $\hat{q}^{\mathsf{approx}}$ is an asymptotically unbiased approximation to $\boldsymbol{q}^{\mathsf{Bayes}}$, the average $L_2$ distance between $\hat{q}^{\mathsf{approx}}$ and $\boldsymbol{q}^{\mathsf{Bayes}}$ is a non-trivial quantity due to the intrinsic variability in $\boldsymbol{X}$. In the following subsection, we introduce our Coordinate wise Adaptive Shrinkage Prediction Rule, CASP, that relies on data driven adaptive shrinkage factors to reduce the marginal variability of $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ for any fixed $\boldsymbol{S}$, and minimize the squared errors of the predictors from $\boldsymbol{q}^{\mathsf{Bayes}}$.

## 3.2 Improved predictive efficiency by coordinate-wise shrinkage

We continue our discussion with respect to the dis-aggregative model ($\boldsymbol{A}=\boldsymbol{I}_n$) and first introduce a class of coordinate-wise shrinkage predictive rules that includes $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ as a special case.

**Definition 2** (Class of coordinate-wise shrinkage predictive rules)**.** *Consider a class of coordinate-wise* *shrinkage predictive rules* $\mathcal{Q}^{\mathsf{cs}} = \{\hat{q}_i^{cs}(\mathbf{X}|\boldsymbol{S}, f_i, \tau, \beta) \mid f_i \in \mathbb{R}_+, \tau \in \mathbf{T}_0, \beta \in \mathbf{B}_0\}$ *where*

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

*with* $\hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S}, \tau, \beta)$ *as defined in definition 1 and* $f_i \in \mathbb{R}_+$ *is a shrinkage factor depending only on* $\boldsymbol{S}$.

The class $\mathcal{Q}^{\mathsf{cs}}$ represents a wider class of predictive rules when compared to the linear functional form of the Bayes rule. In particular, it includes $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ from definition 1 when $f_i = 1$ for all $i$. The coordinate-wise shrinkage factors $f_i$ do not depend on $\boldsymbol{X}$ but only on $\boldsymbol{S}$, and their role lies in reducing the marginal variability of the predictive rule as demonstrated in Lemma 3 below.

**Lemma 3.** *Suppose that assumptions **A1**, **A2** and **A3** hold. Under the hierarchical model of equations* (1) *and* (4)*, as* $n \to \infty$*,*

*(a)* $\mathbb{E}\Big\{\Big(\hat{q}_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\Big)^2\Big\}$ *is minimized at*

$$f_i^{\mathsf{OR}} = \frac{\boldsymbol{e}_i^T U(\boldsymbol{\Sigma})\boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})\hat{H}_{1,-1,\beta}\boldsymbol{e}_i} + O_p\Big(\sqrt{\frac{\log n}{n}}\Big)$$

*where* $U(\boldsymbol{\Sigma}) := H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}$, $\mathcal{J}(\boldsymbol{\Sigma}) := \boldsymbol{\Sigma} + \tau\boldsymbol{\Sigma}^\beta$ *and the expectation is taken with respect to the marginal distribution of* $\boldsymbol{X}$ *with* $\boldsymbol{S}$ *fixed.*

*(b) For any fixed* $\tau$, $\beta$, *with probability 1,*

$$\overline{\lim_{n\to\infty}} \max_{1\le i\le n} f_i^{\mathsf{OR}} \le 1.$$

*Moreover, let* $\mathcal{M} = \{1 \le i \le n : \|\mathbf{P}_K \boldsymbol{e}_i\|_2 > 0\}$, *where* $\mathbf{P}_K$ *denotes the* $K$*-dimensional projection matrix associated with the* $K$ *spiked eigenvalues of* $\boldsymbol{\Sigma}$. *Then, with* $j(x) := x + \tau x^\beta$ *as the scalar version of* $\mathcal{J}(\boldsymbol{\Sigma})$, *we have*

$$\max_{i\in\mathcal{M}} f_i^{\mathsf{OR}} \le \max_{i\in\mathcal{M}} \frac{\boldsymbol{e}_i^T U(\boldsymbol{\Sigma})\boldsymbol{e}_i}{\boldsymbol{e}_i^T U(\boldsymbol{\Sigma})\boldsymbol{e}_i + j(\ell_0)(h_{1,-1,\beta}(\ell_K) - h_{1,-1,\beta}(\ell_0))^2\|\mathbf{P}_K\boldsymbol{e}_i\|_2^2} + O_P\left(\sqrt{\frac{\log n}{n}}\right),$$

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

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

$$\lim_{n\to\infty} \frac{\mathbb{E}\|\hat{\boldsymbol{q}}^{\mathsf{approx}}(\boldsymbol{X}|\boldsymbol{S}, \tau, \beta) - \boldsymbol{q}^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\|_2^2}{\mathbb{E}\|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) - \boldsymbol{q}^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\|_2^2} \ge 1\ ,$$

*where, the expectations are taken with respect to the marginal distribution of* $\boldsymbol{X}$ *with* $\boldsymbol{S}$ *fixed.*

Lemma 3 is proved in the supplementary material. An interesting point to note about the proof of statement $(a)$ of the lemma is that minimizing the squared error essentially reduces to minimizing the variability of $\hat{\boldsymbol{q}}^{\mathsf{cs}}$ as any member in $\mathcal{Q}^{\mathsf{cs}}$ has asymptotically negligible bias. The optimal variance is attained by the oracle shrinkage factors $f_i^{\mathsf{OR}}$ which assume knowledge of $H_{1,-1,\beta}$ and $\mathcal{J}(\boldsymbol{\Sigma})$. Statement $(b)$ shows that these shrinkage factors lie in $[0,1]$. It also shows that some of them are actually quite different from 1. Thus, the resultant coordinate-wise shrunken oracle prediction rule greatly differs from $\hat{\boldsymbol{q}}^{\mathsf{approx}}$. Indeed, statement $(b)$ shows that if the eigenvectors of $\boldsymbol{\Sigma}$ are relatively sparse, so that for a small number of coordinates $i$, the quantities $\|\mathbf{P}_K\boldsymbol{e}_i\|_2$ are positive (and relatively large), then the shrinkage factor $f_i^{OR}$ for the corresponding coordinates can be significantly smaller than 1. Statement $(c)$ trivially follows from $(b)$ and guarantees that $\hat{\boldsymbol{q}}^{\mathsf{cs}}$ constructed based on the oracle shrinkage factors $f_i^{\mathsf{OR}}$ are at least as good as $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ in terms of squared error distance from the true $\boldsymbol{q}^{\mathsf{Bayes}}$ predictor. However, as $\boldsymbol{\Sigma}$ is unknown, $f_i^{\mathsf{OR}}$ cannot be computed in practice. Theorem 1A allows us to estimate the oracle shrinkage factors consistently and those estimates form a key ingredient in our proposed predictive rule CASP in definition 3 below.

**Definition 3** (CASP)**.** *The coordinate-wise adaptive shrinkage prediction rule is given by* $\hat{\boldsymbol{q}}^{\mathsf{casp}} \in \mathcal{Q}^{\mathsf{cs}}$ *with* $f_i = \hat{f}_i^{\mathsf{prop}}$ *where*

$$\hat{f}_i^{\mathsf{prop}} = \frac{\boldsymbol{e}_i^T \tau \hat{H}_{1,\beta-1,\beta}\boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{R}\boldsymbol{e}_i}$$

*and*

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

*with* $j(x) := x + \tau x^{\beta}$ *as the scalar version of* $\mathcal{J}(\boldsymbol{\Sigma})$.

Unlike the numerator, the denominator in $f_i^{\mathsf{OR}}$ is not linear in $H_{r,\alpha,\beta}$ and estimating it with desired precision involves second order terms in $\hat{R}$. Lemma 4 below shows that indeed $\hat{f}_i^{\mathsf{prop}}$ is a consistent estimate of $f_i^{\mathsf{OR}}$ under our hierarchical model.

**Lemma 4.** *Under the hierarchical model of equations* (1) *and* (4)*,*

$$\sup_{1\le i\le n}|\hat{f}_i^{\mathsf{prop}} - f_i^{\mathsf{OR}}| = O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

Using Lemmas 3 (a) and 4, Theorem 2A below guarantees the oracle optimality of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ in the class $\mathcal{Q}^{\mathsf{cs}}$ in the sense that the shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ reduce the squared error between CASP and the Bayes predictive rule as much as the oracle shrinkage factors $f_i^{\mathsf{OR}}$ would for any predictive rule in the class $\mathcal{Q}^{\mathsf{cs}}$.

**Theorem 2A** (Oracle optimality of CASP)**.** *Under assumptions **A1**, **A2** and **A3**, and the hierarchical model of equations* (1) *and* (4)*, we have, conditionally on* $\boldsymbol{X}$*,*

$$\sup_{\tau \in \mathbf{T}_0, \beta \in \mathbf{B}_0} \frac{\|\hat{\boldsymbol{q}}^{\mathsf{casp}}(\boldsymbol{X}|\boldsymbol{S}, \hat{\boldsymbol{f}}^{\mathsf{prop}}, \tau, \beta) - \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta)\|_2^2}{\|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) - \boldsymbol{e}_i^T \boldsymbol{\eta}_0\|_2^2} = O_p(\log n/n) \ .$$

In the following section, we turn our attention to the aggregative model where $\boldsymbol{A} \in \mathbb{R}^{p \times n}$ with $p \leq n$ and $\boldsymbol{A}\boldsymbol{A}^T$ is invertible. The proofs of Lemma 4 and Theorem 2A are provided in the accompanying supplementary material.

# 4 Proposed predictive rule for aggregative predictions

Under the aggregative model, recall that equation (7) expresses $G_{r,\alpha,\beta}$ in terms of $H_{r,\alpha,\beta}$. To estimate $G_{r,\alpha,\beta}$ in this setting, we adopt the substitution principle and construct the following estimates of $G_{r,\alpha,\beta}$

$$\begin{aligned}
\hat{G}_{0,1,0} &= \boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T \\
\hat{G}_{1,0,\beta} &= \tau \boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T\Big[\boldsymbol{A}\Big(\tau\hat{H}_{0,\beta,0} + \hat{H}_{0,1,0}\Big)\boldsymbol{A}^T\Big]^{-1}\boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T \\
\hat{G}_{1,-1,\beta} &= \tau \boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T\Big[\boldsymbol{A}\Big(\tau\hat{H}_{0,\beta,0} + \hat{H}_{0,1,0}\Big)\boldsymbol{A}^T\Big]^{-1}
\end{aligned}$$

which appear in the functional form of CASP for aggregative models in definition 4 below. Throughout this section, we assume the following regularity condition on the aggregation matrix $\boldsymbol{A}$:

**A4 Aggregation matrix:** Suppose $p = o(n)$ and $\boldsymbol{A} \in \mathbb{R}^{p \times n}$ is such that the matrix $\boldsymbol{A}\boldsymbol{A}^T$ is invertible and has uniformly bounded condition number even as $p, n \to \infty$.

**Definition 4** (CASP for aggregative models)**.** *For any fixed* $\boldsymbol{A}$ *obeying assumption **A4**, consider a class of coordinate-wise shrinkage predictive rules* $\mathcal{Q}_{\boldsymbol{A}}^{\mathsf{cs}} = \{\hat{q}_i^{cs}(\boldsymbol{A}\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) \mid f_i \in \mathbb{R}_+, \tau \in \mathbf{T}_0, \beta \in \mathbf{B}_0\}$ *where*

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

*and* $\hat{\mathcal{F}}_i^{\mathsf{loss}}(\boldsymbol{S}, \boldsymbol{A}, \tau, \beta)$ *are the estimates of* $\mathcal{F}_i^{\mathsf{loss}}(\boldsymbol{\Sigma}, \boldsymbol{A}, \tau, \beta)$ *as defined in Lemma 1 with* $G_{r,\alpha,\beta}$ *replaced by* $\hat{G}_{r,\alpha,\beta}$ *and* $f_i \in \mathbb{R}_+$ *are shrinkage factors depending only on* $\boldsymbol{S}$ *and* $\boldsymbol{A}$*. The coordinate-wise adaptive shrinkage predictive rule for the aggregative model is given by* $\hat{\boldsymbol{q}}^{\mathsf{casp}} \in \mathcal{Q}_{\boldsymbol{A}}^{\mathsf{cs}}$ *with* $f_i = \hat{f}_i^{\mathsf{prop}}$ *where*

$$\hat{f}_i^{\mathsf{prop}} = \frac{\boldsymbol{e}_i^T\hat{\mathcal{N}}\boldsymbol{e}_i}{\boldsymbol{e}_i^T\hat{D}\boldsymbol{e}_i}$$

*and*

$$\begin{aligned}\hat{\mathcal{N}} &= \tau \hat{G}_{1,-1,\beta} \boldsymbol{A} \hat{H}_{-1,1+\beta,\beta} \boldsymbol{A}^T \hat{G}_{1,-1,\beta} \\ \hat{D} &= \hat{\mathcal{N}} + j(\hat{\ell}_0^{\mathsf{e}}) \sum_{j=1}^{K} \hat{\zeta}_j^{-4} \Big( h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}}) \Big)^2 \boldsymbol{A} \hat{\mathbf{p}}_j (\boldsymbol{A} \hat{\mathbf{p}}_j)^T\end{aligned}$$

*with $j(x) := x + \tau x^{\beta}$ as the scalar version of $\mathcal{J}(\boldsymbol{\Sigma})$.*

Next, we establish the analogue of Theorem 1A for this set-up. The proof is much more complicated, as for a general $\boldsymbol{A}$, the expression in the posterior covariances loses commutativity in multiplicative operations between $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$. The resultant is that for quadratic form estimation, we need to be precise in tackling the distortion in the spectrum of the posterior variance due to the presence of the linear aggregation matrix $\boldsymbol{A}$. We show that the substitution principle, which avoids higher order corrections, is still consistent under $p \lesssim n^{-1/2}$ situations that hold in most applications in this setting, outside which our consistency bounds deteriorate due to the cost of inversion paid by the simple substitution rules.

**Theorem 1B** (Asymptotic consistency of $\boldsymbol{b}^T \hat{G}_{r,\alpha,\beta} \boldsymbol{b}$)**.** *Under assumptions **A1**, **A2**, **A3** and **A4**, uniformly over $\tau \in \mathbf{T}_0$, $\beta \in \mathbf{B}_0$ 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 $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$*

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

*where the dependence of $G_{r,\alpha,\beta}$ on $\tau$ has been kept implicit for notational ease.*

**Theorem 2B** (Oracle optimality of CASP)**.** *Under assumptions **A1**, **A2**, **A3** and **A4**, and the hierarchical model of equations* (1) *and* (4)*, we have, conditionally on $\boldsymbol{X}$,*

$$\sup_{\tau \in \mathbf{T}_0, \beta \in \mathbf{B}_0} \frac{\left\| \hat{\boldsymbol{q}}^{\mathsf{casp}}(\boldsymbol{A}\boldsymbol{X} | \boldsymbol{S}, \hat{\boldsymbol{f}}^{\mathsf{prop}}, \tau, \beta) - \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{A}\boldsymbol{X} | \boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) \right\|_2^2}{\left\| \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{A}\boldsymbol{X} | \boldsymbol{S}, \boldsymbol{f}^{\mathsf{OR}}, \tau, \beta) - \boldsymbol{e}_i^T \boldsymbol{A} \boldsymbol{\eta}_0 \right\|_2^2} = O_p \Big\{ \max\Big( \Big(\frac{p}{n}\Big)^2, \frac{\log n}{n} \Big) \Big\}$$

Using Theorem 2B, we show that in the aggregative model too the data driven adaptive shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ continue to guarantee the oracle optimality of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ in the class $\mathcal{Q}_{\boldsymbol{A}}^{\mathsf{cs}}$. For the proofs of Theorem 1B and 2B we refer the reader to Appendix A.3 and the accompanying supplementary material, respectively.

**Implementation and R package `casp` -** The R package `casp` has been developed to implement our proposed predictive rule CASP. It uses the following scheme to estimate the prior hyper-parameters $(\tau, \beta)$ and the number of spikes $K$:

For estimating $K$ we use the procedure described in Kritchman and Nadler (2009) that estimates $K$ through a sequence of hypothesis tests determining at each step whether the $k^{th}$ sample eigenvalue came from a spike. To estimate the prior hyper-parameters $(\tau, \beta)$, we first note that marginally $\boldsymbol{X} \sim N_n(\boldsymbol{\eta}_0, \mathcal{J}(\boldsymbol{\Sigma}))$. Let $\mathcal{J}_{\mathsf{inv}}(\boldsymbol{\Sigma}) = (\boldsymbol{\Sigma} + \tau\boldsymbol{\Sigma}^{\beta})^{-1}$. Our scheme for choosing $(\tau, \beta)$ is based on an empirical Bayes approach wherein we maximize the marginal likelihood of $\boldsymbol{X}$ with respect to $(\tau, \beta)$ with $\mathcal{J}(\boldsymbol{\Sigma})$ and $\mathcal{J}_{\mathsf{inv}}(\boldsymbol{\Sigma})$ replaced by their estimates $\hat{\mathcal{J}} = \tau\hat{H}_{-1,1+\beta,\beta}$ and $\hat{\mathcal{J}}_{\mathsf{inv}} = \tau^{-1}\hat{H}_{1,-1-\beta,\beta}$ respectively. In particular an estimate of $(\tau, \beta)$ is given by

$$(\hat{\tau}, \hat{\beta}) = \underset{\tau \in \mathbf{T}_0, \beta \in \mathbf{B}_0}{\arg\max}\ \phi_n\Big\{\big(\boldsymbol{x} - \boldsymbol{\eta}_0\big)^T \hat{\mathcal{J}}_{inv}\big(\boldsymbol{x} - \boldsymbol{\eta}_0\big)\Big\} \tag{14}$$

To facilitate implementation, the minimization in equation (14) is conducted numerically over a bounded interval $[\tau_{\mathsf{lb}}, \tau_{\mathsf{ub}}] \times [\beta_{\mathsf{lb}}, \beta_{\mathsf{ub}}]$ where, in most practical applications, prior knowledge dictates the lower $(\tau_{\mathsf{lb}}, \beta_{\mathsf{lb}})$ and upper bounds $(\tau_{\mathsf{ub}}, \beta_{\mathsf{ub}})$ of the above intervals. For example, in finance applications (Kozak et al., 2017), it is usually known if $\tau < 1$ and $\beta \geq 1$. In light of this prior knowledge, our implementation conducts the minimization on $10 \times 10$ equi-spaced points over the bivariate grid $[0.01,\ 1] \times [1,\ 5]$. Similarly, if it is known that $\tau \geq 1$ and $\beta < 1$, the minimization is conducted on $10 \times 10$ equi-spaced points over the bivariate grid $[1,\ 5] \times [0.1,\ 1]$. The R-package implementation also allows problem specific lower and upper bounds and a choice on the coarseness of the grid. In the simulations and real data examples of sections 5 and 6, we use the above scheme to estimate $(\tau,\ \beta)$.

# 5 Simulation Studies

In this section we asses the predictive performance of CASP across a wide range of simulation experiments. We consider four competing predictive rules that use different methodologies to estimate $\boldsymbol{\Sigma}$ and thereafter plug-in their respective estimates of $\boldsymbol{\Sigma}$ in the Bayes predictive rule of Lemma 1. In what follows, we briefly discuss these competing methods for estimating $\boldsymbol{\Sigma}$:

1. $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ - the predictive rule that uses the bi-cross-validation approach of Owen and Wang (2016) which, under a heteroscedastic factor model structure, first estimates the number of factors, then constructs an estimate $\boldsymbol{S}^{\mathsf{Bcv}}$ of $\boldsymbol{\Sigma}$ and finally plugs-in $\boldsymbol{S}^{\mathsf{Bcv}}$ in the Bayes predictive rule of Lemma 1. We use the implementation available in the R package `esaBcv` for our simulations.

2. $\hat{\boldsymbol{q}}^{\mathsf{Fact}}$ - the predictive rule that uses the FactMLE algorithm of Khamaru and Mazumder (2018)

to estimate $\boldsymbol{S}^{\mathsf{Fact}}$ by formulating the low rank maximum likelihood Factor Analysis problem as a non-linear, non-smooth semidefinite optimization problem. The implementation of the FactMLE algorithm is available in the R package `FACTMLE` wherein we use an estimate $\hat{K}$ of $K$ as discussed in section 4.

3. $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ - the predictive rule that uses the approach of Fan et al. (2013) to estimate $\boldsymbol{S}^{\mathsf{Poet}}$ by first retaining the first $\hat{K}$ principal components of $\boldsymbol{S}$ and then using a thresholding procedure on the remaining sample covariance matrix $\boldsymbol{S}$. The implementation of this approach is available in the R-package `POET` where $\hat{K}$ is an estimate of the number of spikes from section 4.

4. $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$ - the Naive predictive rule which first estimates the number of spikes $\hat{K}$ from the data, reconstructs the sample covariance matrix $\boldsymbol{S}^{\mathsf{Naive}}$ from the leading $\hat{K}$ eigen values and eigen vectors of $\boldsymbol{S}$, and finally plugs in $\boldsymbol{S}^{\mathsf{Naive}}$ in place of $\boldsymbol{\Sigma}$ in the Bayes predictive rule $\boldsymbol{q}^{\mathsf{Bayes}}$ in Lemma 1.

To assess the performance of various predictive rules, we calculate a relative estimation error (REE) which is defined as

$$\text{REE}(\hat{\boldsymbol{q}}) = \frac{R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}) - R_n(\boldsymbol{\theta}, \boldsymbol{q}^{\mathsf{Bayes}})}{R_n(\boldsymbol{\theta}, \hat{\boldsymbol{q}}^{\mathsf{approx}}) - R_n(\boldsymbol{\theta}, \boldsymbol{q}^{\mathsf{Bayes}})}$$

where $\hat{\boldsymbol{q}}$ is any prediction rule, $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ is CASP with shrinkage factors $f_i = 1$ for all $i$ and $\boldsymbol{q}^{\mathsf{Bayes}}$ is the Bayes predictive rule based on the knowledge of unknown $\boldsymbol{\Sigma}$. A value of REE larger than 1 implies poorer prediction performance of $\hat{\boldsymbol{q}}$ relative to $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ whereas a value smaller than 1 implies a better prediction performance. In particular, REE allows us to quantify the relative advantage of using coordinate wise adaptive shrinkage in our proposed predictive rule $\hat{\boldsymbol{q}}^{\mathsf{casp}}$.

### 5.1 Experiment 1

In the setup of experiment 1, we consider prediction under the dis-aggregated model and 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 impose a spike covariance structure on $\boldsymbol{\Sigma}$ with $K = 10$ spikes under the following two scenarios with $l_0$ fixed at 1.

- Scenario 1: we consider the generalized absolute loss function in equation (5) with $b_i$ sampled uniformly between $(0.9, 0.95)$, $h_i = 1 - b_i$ with $(\tau, \beta) = (0.5, 0.25)$ and $K$ spikes equi-spaced between 80 and 20.

- Scenario 2: we consider the linex loss function in equation (6) with $a_i$ sampled uniformly between $(-2, -1)$, $b_i = 1$ with $(\tau, \beta) = (0.01, 2)$ and $K$ spikes equi-spaced between 25 and 10.

To estimate $\boldsymbol{S}$, we sample $\boldsymbol{W}_j$ from $N_n(\mathbf{0}, \boldsymbol{\Sigma})$ independently for $m$ samples where we allow $m$ to vary over $(15, 20, 25, 30, 35, 40, 45, 50)$. Finally $m_x = 1$ copy of $\boldsymbol{X}$ is sampled from $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with $m_0 = 1$.

This sampling scheme is repeated over $1,000$ repetitions and the average REE of the competing predictive rules and CASP is presented in figures 1 and 2 for scenarios 1 and 2 respectively. In tables 1 and 2, we report the average REE at $m = 15$. Using the R-package `POET`, the estimation of $\boldsymbol{S}^{\mathsf{Poet}}$ was extremely slow in our simulations and therefore we report the average REE of $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ only at $m = 15$ and exclude this predictive rule from the figures.

**Table 1:** Scenario 1: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ and averaged over $1,000$ repetitions

| | $\hat{K}$ | $\hat{\tau}$ | $\hat{\beta}$ | REE |
|---|---|---|---|---|
| CASP | 7.5 | 0.92 | 0.37 | **0.948** |
| Bcv | 3.4 | 0.91 | 0.27 | 1.144 |
| FactMLE | 7.5 | 0.92 | 0.15 | 1.594 |
| POET | 7.5 | 0.92 | 0.15 | 1.959 |
| Naïve | 7.5 | 0.97 | 0.23 | 1.353 |

**Table 2:** Scenario 2: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ and averaged over $1,000$ repetitions

| | $\hat{K}$ | $\hat{\tau}$ | $\hat{\beta}$ | REE |
|---|---|---|---|---|
| CASP | 4.7 | 0.27 | 1.23 | **0.999** |
| Bcv | 0.6 | 0.34 | 1.1 | 1.046 |
| FactMLE | 4.7 | 0.66 | 1.09 | 1.475 |
| POET | 4.7 | 0.66 | 1.08 | 1.312 |
| Naïve | 4.7 | 0.64 | 1.15 | 1.333 |

The left panels of figures 1 and 2 both suggest a superior risk performance of CASP as $m$ varies. Moreover, when the ratio $n/(m-1)$ is largest, the right panels of these figures plot the sorted shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ averaged over the $1,000$ repetitions (red line) and sandwiched between its $10^{th}$ and $90^{th}$ percentiles (represented by the gray shaded region) under the two scenarios. Under scenario 1 in particular, the estimated shrinkage factors are all smaller than $1$ indicating the significant role that the coordinate-wise shrinkage plays in reducing the marginal mean square error of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ from $\boldsymbol{q}^{\mathsf{Bayes}}$. However as $\beta$ increases from $0.25$ to $2$ in scenario 2, the estimated shrinkage factors move closer to $1$, and the risk performances of $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ and $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ are indistinguishable from each other as seen in table 2 wherein the REE of CASP is almost $1$. This is not unexpected because with a fixed $\tau > 0$ and $\beta$ growing above $1$, the factor $\sum_{j=1}^{K}\hat{\zeta}_j^{-4}\Big(h_{1,-1,\beta}(\hat{\ell}_j^{\mathsf{e}}) - h_{1,-1,\beta}(\hat{\ell}_0^{\mathsf{e}})\Big)^2$ 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 tables 1 and 2, $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ is the most competitive predictive rule next to $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ 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.

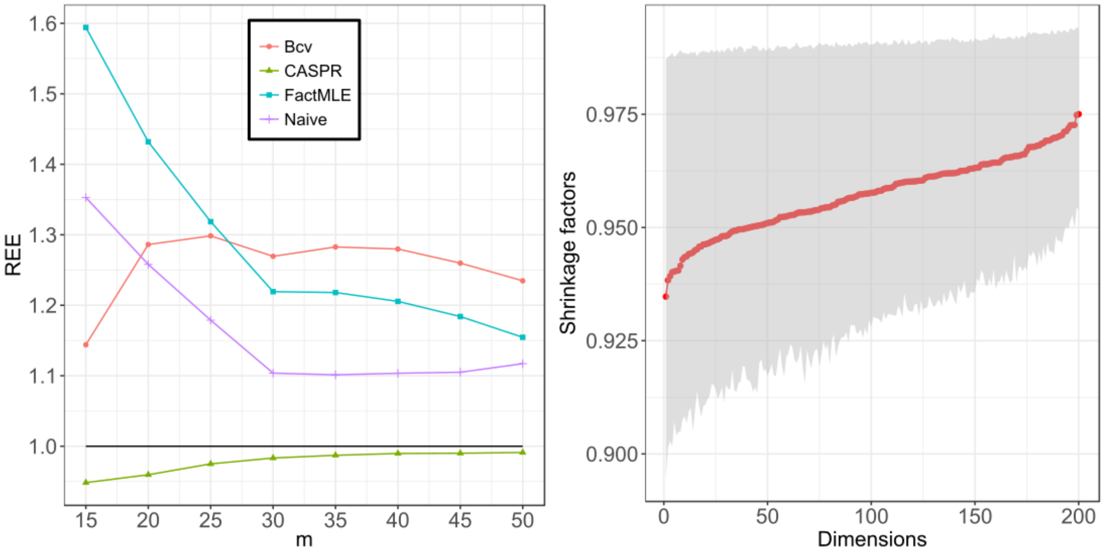

**Figure 1:** Experiment 1 Scenario 1 (Generalized absolute loss): Left - Relative Error estimates as $m$ varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over $1,000$ repetitions at $m = 15$ and sandwiched between its $10^{th}$ and $90^{th}$ percentiles

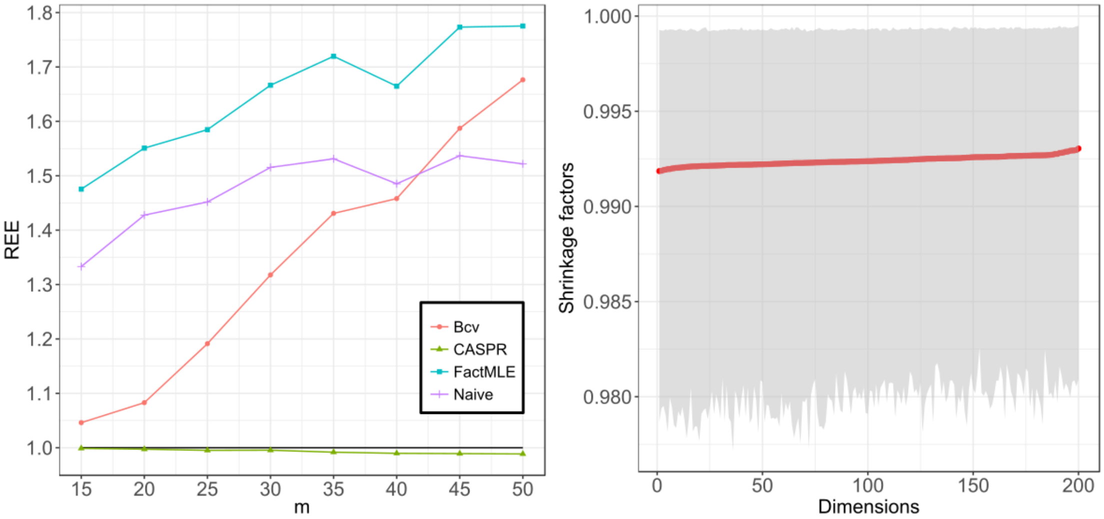

**Figure 2:** Experiment 1 Scenario 2 (Linex loss): Left - Relative Error estimates as $m$ varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over $1,000$ repetitions at $m = 15$ and sandwiched between its $10^{th}$ and $90^{th}$ percentiles

## 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:

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

where $\boldsymbol{B}$ is the $n \times K$ matrix of factor loadings, $\Gamma_t$ is the $K \times 1$ vector of latent factors, $K \ll n$ represents the number of latent factors 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 Fan et al. (2013), Khamaru and Mazumder (2018), Owen and Wang (2016) for estimating $\boldsymbol{\Sigma}$. Thus the three competing predictive rules $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$, $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$ and $\hat{\boldsymbol{q}}^{\mathsf{Fact}}$ are well suited for prediction in this model. Factor models of this form are often considered in portfolio risk estimation (see for example Fan et al. (2016)) where the goal is to first estimate the matrix of factor loadings $\boldsymbol{B}$ and the vector of latent factors $\boldsymbol{\Gamma}_t$ and thereafter use the fitted model to sequentially predict $\boldsymbol{A}\boldsymbol{X}_{t+s}$ for $s = 1, 2, \ldots, T$ where $\boldsymbol{X}_t$ might represent an $n$ dimensional vector of stock excess returns and $\boldsymbol{A}$ is the $p \times n$ weight matrix that aggregates the predicted excess returns into $p \ll n$ individual portfolios level returns. Often an autoregressive structure is imposed on $\boldsymbol{\Gamma}_t$ so that $\boldsymbol{\Gamma}_t = \boldsymbol{\Phi}\boldsymbol{\Gamma}_{t-1} + \boldsymbol{v}_t$ which is the so called dynamic factor model (Geweke, 1977) where $\boldsymbol{\Phi}$ is a $K \times K$ matrix of autoregressive coefficients and $\boldsymbol{v}_t \sim N_K(\boldsymbol{0}, \boldsymbol{D})$.

For the purposes of this simulation exercise we take $\boldsymbol{\Phi} = \boldsymbol{0}$ and $\boldsymbol{D} = \boldsymbol{I}_K$ with $K = 10$ factors. We fix $\boldsymbol{\eta}_0 = 0$ and simulate the rows of $\boldsymbol{B}$ from $N_K(\boldsymbol{0}, c\boldsymbol{I}_K)$. The elements of the aggregation matrix $\boldsymbol{A}$ are simulated uniformly from $(0, 1)$ with $p = 20$ rows normalized to 1. In this experiment, similar to experiment 1, we consider two scenarios:

- Scenario 1: we fix $(c, \tau, \beta) = (0.5, 0.5, 0.25)$ and simulate the diagonal elements of $\boldsymbol{\Delta}_n$ uniformly from $(0.25, 1.5)$.
- Scenario 2: we fix $(c, \tau, \beta) = (0.25, 0.01, 2)$ and simulate the diagonal elements of $\boldsymbol{\Delta}_n$ uniformly from $(1, 1.5)$.

To estimate $\boldsymbol{S}$, we sample $\boldsymbol{W}_j$ from $N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ independently for $m$ samples where we allow $m$ to vary over $(15, 20, 25, 30, 35, 40, 45, 50)$. Finally $m_x = 1$ copy of $\boldsymbol{X}_t$ is sampled from $N_n(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and the goal is to predict $\boldsymbol{A}\boldsymbol{X}_{t+1}$ under a linex loss with $b_i = 1$ and $a_i \overset{i.i.d}{\sim} \mathsf{Unif}(1, 2)$ to emphasize the severity of over prediction of portfolio excess returns. This sampling scheme is repeated over $1,000$ repetitions and the average REE of the competing predictive rules and CASP is presented in figures 3 and 4 for scenarios 1 and 2 respectively. We see a similar phenomenon as in experiment 1 where the relative error estimates of CASP suggest superior risk performance in both the scenarios. From tables 3 and 4, $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ is still the most competitive predictive rule after CASP but continues to under-estimate the number of factors especially in scenario 2. The estimated shrinkage factors $f_i,\ i = 1, \cdots, p$, in the right panel of figure 3 exhibits magnitudes smaller than 1 which explains the relatively smaller REE of CASP in scenario 1. This trend is, however, reversed in scenario 2 where $\beta > 1$ and CASP is only marginally better than $\hat{\boldsymbol{q}}^{\mathsf{approx}}$.

**Table 3:** Experiment 2 Scenario 1: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ and averaged over $1,000$ repetitions

| | $\hat{K}$ | $\hat{\tau}$ | $\hat{\beta}$ | REE |
|---|---|---|---|---|
| CASP | 7.8 | 0.94 | 0.34 | **0.909** |
| Bcv | 3.8 | 0.92 | 0.26 | 1.014 |
| FactMLE | 7.8 | 0.92 | 0.15 | 1.390 |
| POET | 7.8 | 0.92 | 0.15 | 1.671 |
| Naïve | 7.8 | 0.95 | 0.20 | 1.463 |

**Table 4:** Experiment 2 Scenario 2: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ and averaged over $1,000$ repetitions

| | $\hat{K}$ | $\hat{\tau}$ | $\hat{\beta}$ | REE |
|---|---|---|---|---|
| CASP | 2.5 | 0.12 | 1.29 | **0.986** |
| Bcv | 0.19 | 0.25 | 1.10 | 2.225 |
| FactMLE | 2.5 | 0.52 | 1.07 | 5.654 |
| POET | 2.5 | 0.66 | 1.08 | 7.216 |
| Naïve | 2.5 | 0.36 | 1.14 | 3.636 |

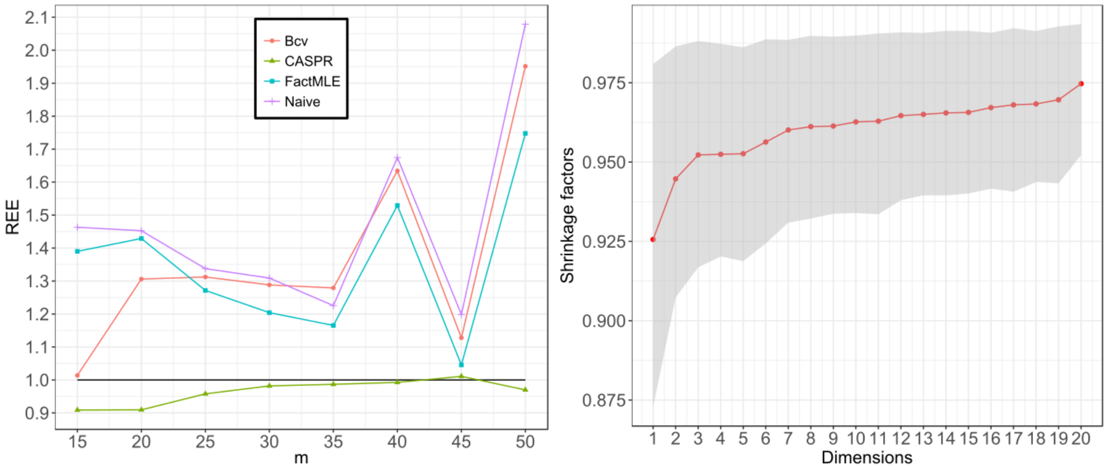

**Figure 3:** Experiment 2 Scenario 1 (Linex loss): Left - Relative Error estimates as $m$ varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over $1,000$ repetitions at $m = 15$ and sandwiched between its $10^{th}$ and $90^{th}$ percentiles

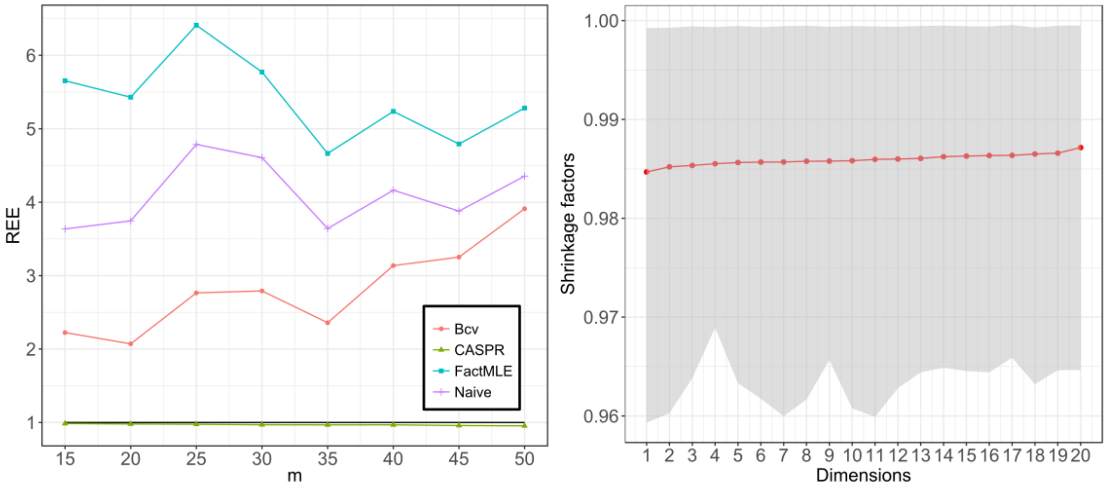

**Figure 4:** Experiment 2 Scenario 2 (Linex loss): Left - Relative Error estimates as $m$ varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over $1,000$ repetitions at $m = 15$ and sandwiched between its $10^{th}$ and $90^{th}$ percentiles

### 5.3 Experiment 3

For experiment 3, we consider a slightly different setup where we do not impose a spike covariance structure on $\mathbf{\Sigma}$. Instead, we assume that $(\mathbf{\Sigma})_{ij} = Cov(X_i, X_j) = 0.9^{|i-j|}$ 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 $\tau\mathbf{\Sigma}^\beta$. We vary $(\tau, \beta)$ across two scenarios where we take $(\tau, \beta)$ 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 $N_n(\boldsymbol{\theta}, \mathbf{\Sigma})$ 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 $\boldsymbol{Y} \sim N_n(\boldsymbol{\theta}, \mathbf{\Sigma})$ 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 $0.9\delta_0 + (1 - 0.9)\mathsf{Unif}(0, 1)$ and normalized to 1 thereafter. This sampling scheme is repeated over $1,000$ repetitions and the average REE of the competing predictive rules and CASP is presented in figures 5, 6 and tables 5, 6 for scenarios 1 and 2 respectively.

**Table 5:** Experiment 3 Scenario 1: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ and averaged over $1,000$ repetitions

| | $\hat{K}$ | $\hat{\tau}$ | $\hat{\beta}$ | REE |
|---|---|---|---|---|
| CASP | 7.5 | 0.87 | 0.22 | **0.919** |
| Bcv | 1.5 | 0.80 | 0.19 | 2.471 |
| FactMLE | 7.5 | 0.91 | 0.15 | 1.271 |
| POET | 7.5 | 0.91 | 0.15 | 1.594 |
| Naïve | 7.5 | 0.91 | 0.15 | 1.306 |

**Table 6:** Experiment 3 Scenario 2: Relative Error estimates (REE) of the competing predictive rules at $m = 15$ and averaged over $1,000$ repetitions

| | $\hat{K}$ | $\hat{\tau}$ | $\hat{\beta}$ | REE |
|---|---|---|---|---|
| CASP | 7.5 | 0.91 | 1.10 | **1.000** |
| Bcv | 1.5 | 0.96 | 1.17 | 1.732 |
| FactMLE | 7.5 | 0.52 | 1.07 | 1.329 |
| POET | 7.5 | 0.85 | 1.05 | 1.505 |
| Naïve | 7.5 | 0.85 | 1.05 | 1.082 |

In this setup, the relative estimation error of CASP continues to be the smallest amongst all the other competing rules, however, for scenario 2 $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ and $\hat{\boldsymbol{q}}^{\mathsf{approx}}$ are almost identical in their performance. Amongst the competing methods $\hat{\boldsymbol{q}}^{\mathsf{Bcv}}$ has the poorest performance 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. On the other hand, $\hat{\boldsymbol{q}}^{\mathsf{Fact}}$ and $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$ exhibit relatively better REE than those seen in experiments 1 and 2 but maintain higher relative estimation errors than $\hat{\boldsymbol{q}}^{\mathsf{casp}}$ indicating potential robustness of CASP to mis-specifications 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

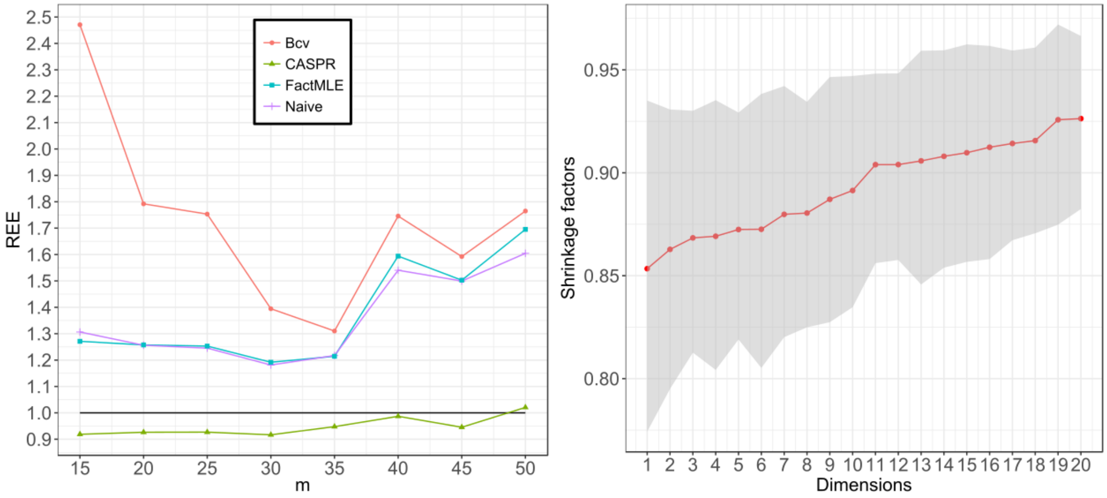

**Figure 5:** Experiment 3 Scenario 1 (Generalized absolute loss): Left - Relative Error estimates as $m$ varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over $1,000$ repetitions at $m = 15$ and sandwiched between its $10^{th}$ and $90^{th}$ percentiles

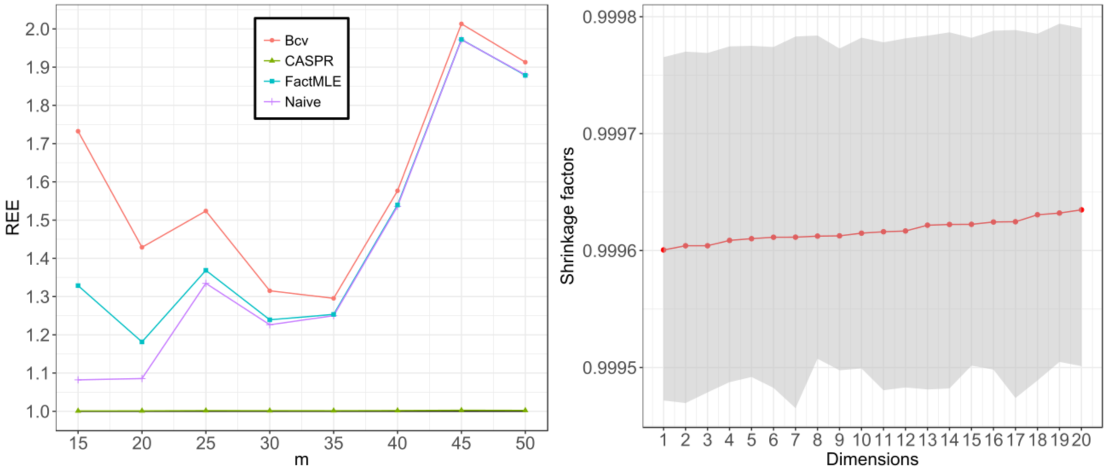

**Figure 6:** Experiment 3 Scenario 2 (Generalized absolute loss): Left - Relative Error estimates as $m$ varies over $(15, 20, 25, 30, 35, 40, 45, 50)$. Right: Magnitude of the sorted shrinkage factors $\hat{f}_i^{\text{prop}}$ averaged over $1,000$ repetitions at $m = 15$ and sandwiched between its $10^{th}$ and $90^{th}$ percentiles

(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 dis-aggregate 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 (2), 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 $t = 1, \ldots, T$. 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 merger and even closure during the prediction period which is often recorded as 0 product sales. Let $\mathbb{X}_0^{(p)}$ be the $n$ dimensional vector denoting the number of units of product $p$ sold across the $n$ outlets in week 0 - October 31, 2011 to November 6, 2011. For our prediction problem, we use a threshold of $s_p$ units for product $p$ and consider only those outlets that have sold atleast $s_p$ units in week 0.

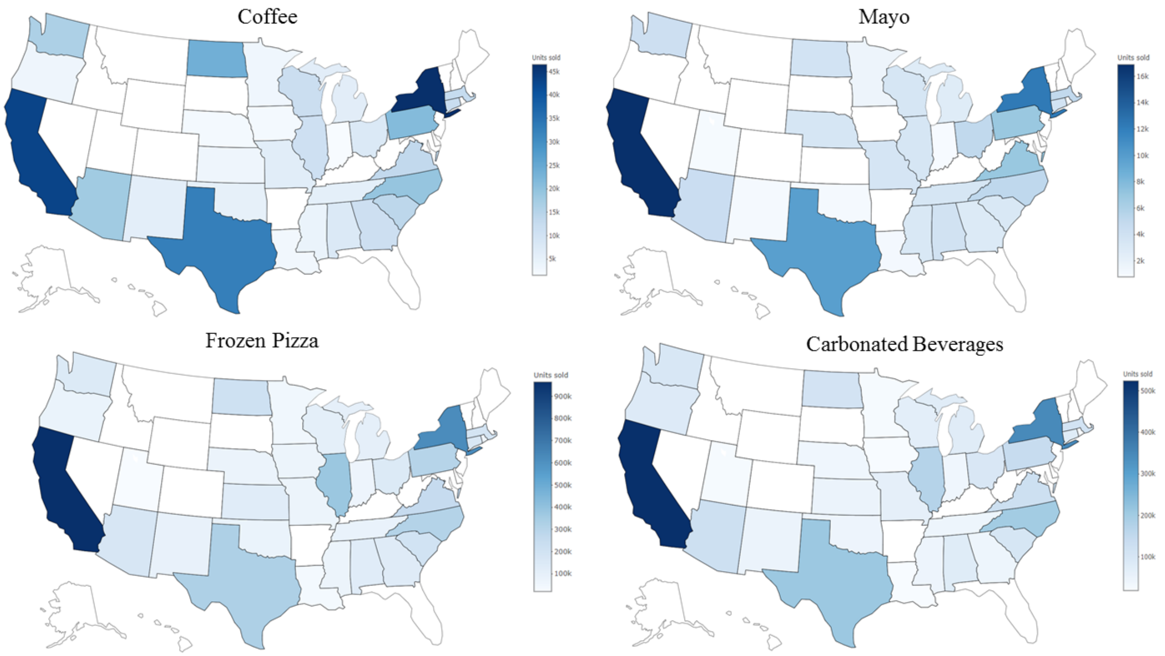

**Figure 7:** CASP predicted weekly demand of the grocery items across US averaged over the two prediction weeks - November 7, 2011 to November 20, 2011.

Let $\boldsymbol{X}_{t-1}^{(p)}$ be the $n_p = \sum_{i=1}^{n} I(\mathbb{X}_{0,i}^{(p)} \geq s_p)$ dimensional vector denoting the number of units of product $p$ sold across $n_p$ stores in week $t-1$. For a distributor, it is economically important to predict the aggregated demands (future sales) for each US state as intra-state transport of inventories and transfer of business and tax accounts can be easily executed within state boundaries. The time $t-1$ prediction problem then is to predict $\boldsymbol{V}_t^{(p)} = \boldsymbol{A}^{(p)}\boldsymbol{X}_t^{(p)}$ where $\boldsymbol{A}^{(p)}$ is a $d_p \times n_p$ matrix that aggregates product $p$ sales across $n_p$ stores into $d_p$ unique states across the U.S. To evaluate the performance of any predictive rule $\hat{\boldsymbol{q}}_t^{(p)}$, we use the generalized absolute loss function of equation (5) and calculate the time $t$ ratio of

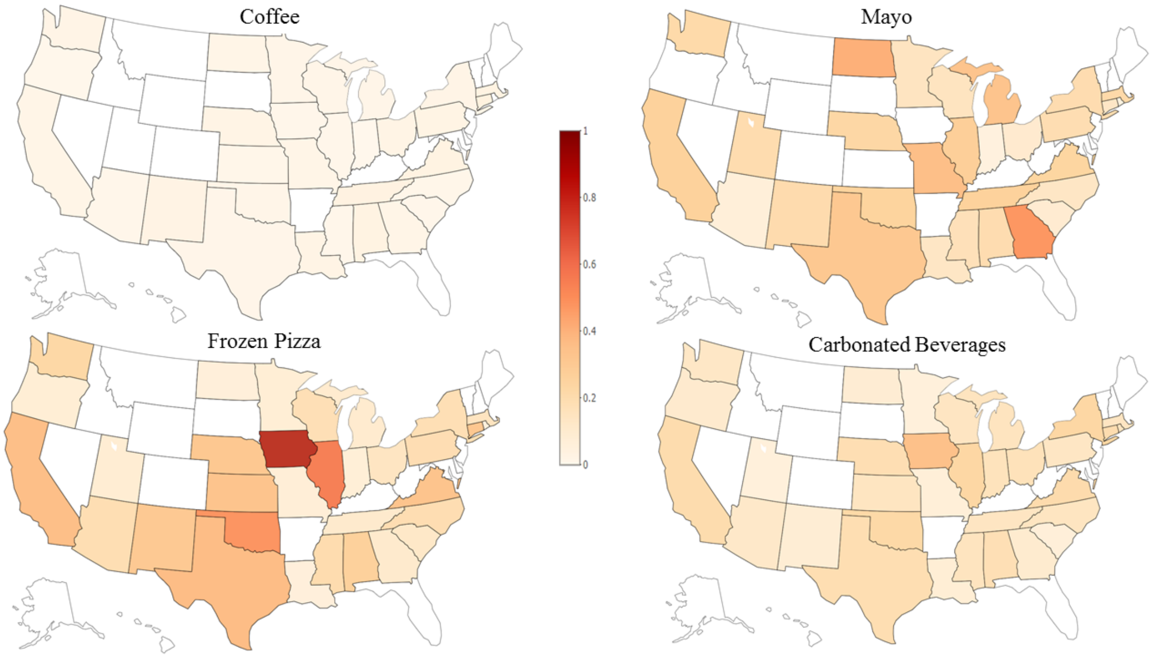

**Figure 8:** Role of coordinate-wise shrinkage in CASP across US states for four grocery items. In the figures, $1-$ the shrinkage factors is displayed and so, deeper shades denote higher shrinkage.

total loss for prediction using $\hat{\boldsymbol{q}}_t^{(p)}$ to the total loss for prediction using CASP with all shrinkage factors $f_i = 1$:

$$\mathcal{L}_t\left(\boldsymbol{q}_t^{\text{(approx,p)}}, \boldsymbol{q}_t^{(p)}\right) = \frac{\sum_{i=1}^{d_p}\left\{b_i\left(V_{t,i}^{(p)} - \hat{q}_{t,i}^{(p)}\right)^+ + h_i\left(\hat{q}_{t,i}^{(p)} - V_{t,i}^{(p)}\right)^+\right\}}{\sum_{i=1}^{d_p}\left\{b_i\left(V_{t,i}^{(p)} - \hat{q}_{t,i}^{\text{(approx,p)}}\right)^+ + h_i\left(\hat{q}_{t,i}^{\text{(approx,p)}} - V_{t,i}^{(p)}\right)^+\right\}} \tag{15}$$

where $\hat{\boldsymbol{q}}_t^{\text{(approx,p)}}$ is CASP with $f_i = 1$ for all $i$. We set $b_i = 0.95$ and $h_i = 1 - b_i$ for all $i = 1, \ldots, d_p$ to emphasize the severity of under-prediction since over-stocking may lead to holding and storage costs as all of these four products have considerably longer expiration dates but under-stocking, on the other hand, may translate into substantial lost sales and reputation costs for the retail outlets. In table 7 we report this loss ratio $\mathcal{L}_t$ for each product $p$ in columns $(a)$ and $(b)$, and for six competing predictive rules: (i) CASP, (ii) the Naive predictive rule as discussed in section 5, (iii) Bcv (Owen and Wang, 2016), (iv) POET (Fan et al., 2013), (v) FactMLE (Khamaru and Mazumder, 2018), and (vi) the Unshrunk predictive rule that simply uses past week's sales to predict the sales in the forthcoming week. To compute an estimate $\boldsymbol{S}^{(p)}$ of the $n_p \times n_p$ population covariance matrix $\boldsymbol{\Sigma}^{(p)}$ of $\boldsymbol{X}_t^{(p)}$ we rely on the additional data on $m = 100$ weeks available from December 31, 2007 to November 29, 2009 and estimate $\boldsymbol{S}^{(p)}$ using the methodology described in Section 2.3. In particular we use the function `smooth.spline` from the R-package `splines2` and choose $k = 3$ knots corresponding to the $25, 50$ and $95$ percentiles of the sales distribution across the $n_p$ stores at each of the $m$ weeks. We complete the specification of our model by setting $\boldsymbol{\eta} = \eta_0^{(p)}\mathbf{1}$ where $\eta_0^{(p)}$ is the median of average weekly sales of $n_p$ outlets over the $m$ weeks and use equation (14) to estimate $\beta^{(p)}$ over the interval $[0.1, 1]$ with $\tau^{(p)}$ fixed at 1.

**Table 7:** Loss ratios (15) across six predictive rules for four products.

| Product | Method | K | $(a.)$ Total sales by state<br>Week 1 Loss Ratio | $(b.)$ Total sales by state<br>Week 2 Loss Ratio |
|---|---|---|---|---|
| Coffee<br>$s_p = 1000,\ n_p = 233,\ d_p = 31$ | CASP | 26 | **0.999** | **1.002** |
| | Naïve | 26 | 1.044 | 1.063 |
| | Bcv | 17 | 1.043 | 1.036 |
| | POET | 26 | 1.047 | 1.070 |
| | FactMLE | 26 | 1.009 | 1.044 |
| | Unshrunk | - | 1.838 | 2.273 |
| Mayo<br>$s_p = 500,\ n_p = 157,\ d_p = 30$ | CASP | 26 | **0.995** | **1.004** |
| | Naïve | 26 | 0.996 | 1.016 |
| | Bcv | 19 | 1.040 | 1.019 |
| | POET | 26 | 0.996 | 1.022 |
| | FactMLE | 26 | 0.999 | 1.012 |
| | Unshrunk | - | 1.084 | 2.420 |
| Frozen Pizza<br>$s_p = 1000,\ n_p = 359,\ d_p = 33$ | CASP | 33 | **1.000** | **0.998** |
| | Naïve | 33 | 1.177 | 1.135 |
| | Bcv | 19 | 1.059 | 1.091 |
| | POET | 33 | 1.033 | 1.040 |
| | FactMLE | 33 | 1.008 | 1.020 |
| | Unshrunk | - | 4.424 | 6.701 |
| Carb. Beverages<br>$s_p = 5000,\ n_p = 410,\ d_p = 33$ | CASP | 37 | **1.003** | **0.984** |
| | Naïve | 37 | 1.065 | 1.033 |
| | Bcv | 20 | 1.073 | 1.142 |
| | POET | 37 | 1.065 | 1.038 |
| | FactMLE | 37 | 1.067 | 1.059 |
| | Unshrunk | - | 3.459 | 8.885 |

The loss ratios reported in columns $(a)$ and $(b)$ of table 7 indicate a competitive performance of CASP over the five remaining predictive rules. CASP continues to provide the smallest loss ratios across both the weeks with the only exception being the loss ratio in Week 1 (column $(a)$) for product 'Mayo', where CASP is competitive with the predictive rules $\hat{\boldsymbol{q}}^{\mathsf{Naive}}$(Naive) and $\hat{\boldsymbol{q}}^{\mathsf{Poet}}$(POET). It is interesting to note that in atleast one of the two weeks, the loss ratio of CASP is marginally bigger than 1 across all four product categories, indicating that the coordiante-wise shrinkage factors do not always bring in any significant improvement in prediction performance. This is not entirely unexpected because the hierarchical model assumption of equation (4) may not hold in this setting and thus the model based shrinkage factors $\hat{f}_i^{\mathsf{prop}}$ may not be the most optimal coordinate-wise shrinkage. Figure 7 presents the CASP predicted future demand across the different states while figure 8 presents the state-wise distribution of the shrinkage factors of CASP for the four products and are plotted as $1 - f_i$ so that a lighter shade in the heatmap corresponds to a smaller shrinkage (larger $\hat{f}_i^{\mathsf{prop}}$). For example in the case of Coffee, the shrinkage factors are all closer to 1 across all the $d_p = 31$ states and this effect translates into loss ratios being almost equal to 1 across the two weeks in table 7. In the case of Frozen Pizza, however, the magnitudes of the shrinkage factors are evenly distributed across the $d_p = 33$ states. For instance, the states of Iowa followed by Illinois and Oklahoma exhibit the largest shrinkages while for Mayo, Georgia and North Dakota have the largest shrinkages in their predicted weekly total sales. In particular, this demonstrates that the shrinkage factors vary across the products because the variability in sales is product specific. More importantly, CASP is flexible enough to capture this inter-product differences and mainly due to its bias correction step (see section 3.1), CASP offers better estimates of future sales than the other popular predictive approaches considered here.

# A Proofs

Due to shortage of space we present the detailed proofs of Theorem 1A and 1B here and the proofs of all the other results in this paper are provided in the supplementary materials.

## A.1 Preliminary expansions for eigenvector and eigenvalues

In this subsection, we put together the key expansions that are needed to prove the theorems. We first express the $j$-th sample eigenvector $\hat{\mathbf{p}}_j$ as

$$\hat{\mathbf{p}}_j = a_j \mathbf{P}_K(\mathbf{e}_{j,K} + \delta_j) + \sqrt{1 - a_j^2}\mathbf{P}_{K,\perp}\mathbf{u}_{j,n-K}, \tag{16}$$

where $\mathbf{P}_K = [\mathbf{p}_1 : \cdots : \mathbf{p}_K]$, $\mathbf{P}_{K,\perp}$ is an $n \times (n-K)$ matrix so that $[\mathbf{P}_K : \mathbf{P}_{K,\perp}]$ is an orthogonal matrix, $a_j = \|\mathbf{P}_K \hat{\mathbf{p}}_j\| \in (0,1)$ (without loss of generality, choosing the correct sign), $\mathbf{e}_{j,K}$ is the $j$-th canonical coordinate vector in $\mathbb{R}^K$. Moreover, $\mathbf{u}_{j,n-K}$ is uniformly distributed on $\mathbb{S}_{n-K-1}$ (the unit sphere in $\mathbb{R}^{n-K}$), so that $\mathbf{u}_{j,n-K} = \varepsilon_j/\sqrt{n-K}$ where $\varepsilon_j \sim N(0, I_{n-K})$. We shall make use of the following asymptotic expansions (Paul, 2007).

$$\|\delta_j\| = O_P(n^{-1/2}) \qquad \text{and} \qquad a_j = \zeta_j + O_P(n^{-1/2}) \tag{17}$$

Now, for $p \le n$, let $\mathbf{A}$ be any $p \times n$ matrix such that $\|\mathbf{A}\|$ and $\|(\mathbf{A}\mathbf{A}^T)^{-1}\|$ are bounded even as $p, n \to \infty$. Then, for any $\mathbf{b} \in \mathbb{R}^p$ with $\|\mathbf{b}\|_2 = 1$, we have the expansion

$$\begin{aligned}\langle \mathbf{b}, \mathbf{A}\hat{\mathbf{p}}_j \rangle \quad &= \quad \zeta_j \langle \mathbf{b}, \mathbf{A}\mathbf{p}_j \rangle + \frac{\sqrt{1-\zeta_j^2}}{\sqrt{n-K}} \langle \mathbf{b}, \mathbf{A}\mathbf{P}_{K,\perp}\varepsilon_j \rangle \\ &\quad + (a_j - \zeta_j)\langle \mathbf{b}, \mathbf{A}\mathbf{p}_j \rangle + (\sqrt{1-a_j^2} - \sqrt{1-\zeta_j^2})\frac{1}{\sqrt{n-K}}\langle \mathbf{b}, \mathbf{A}\mathbf{P}_{K,\perp}\varepsilon_j \rangle \\ &\quad + a_j \langle \mathbf{b}, \mathbf{A}\mathbf{P}_K \delta_j \rangle + \sqrt{1-a_j^2}\langle \mathbf{b}, \mathbf{A}\mathbf{P}_{K,\perp}\varepsilon_j \rangle (\|\varepsilon_j\|^{-1} - (n-K)^{-1/2}).\end{aligned} \tag{18}$$

Suppose that $\mathcal{B}$ be any collection of unit vectors in $\mathbb{R}^p$ of cardinality $O(n^c)$ for some fixed $c \in (0, \infty)$. Then, from (18) we conclude that, uniformly over $\mathbf{b} \in \mathcal{B}$,

$$\langle \mathbf{b}, \mathbf{A}\hat{\mathbf{p}}_j \rangle - \zeta_j \langle \mathbf{b}, \mathbf{A}\mathbf{p}_j \rangle = O_P(\sqrt{\log n/n}) \tag{19}$$

Here, we used the fact that $\langle \mathbf{b}, \mathbf{A}\mathbf{P}_{K,\perp}\varepsilon_j \rangle \sim N(0, \mathbf{b}^T\mathbf{A}(\mathrm{I} - \mathbf{P}_K\mathbf{P}_K^T)\mathbf{A}^T\mathbf{b})$, $|\langle \mathbf{b}, \mathbf{A}\mathbf{p}_j \rangle| \leq \|\mathbf{A}\|$ and, $|\langle \mathbf{b}, \mathbf{A}\delta_j \rangle| \leq \|\mathbf{A}\|\|\delta_j\| = O_P(n^{-1/2})$. Moreover, $|a_j - \zeta_j| = O_P(n^{-1/2})$ implies $|\sqrt{1 - a_j^2} - \sqrt{1 - \zeta_j^2}| = O_P(n^{-1/2})$ and $|\|\varepsilon_j\|^{-1} - (n - K)^{-1/2}| = O_P(n^{-1})$.

## A.2 Proof of Theorem 1A

First note that for any fixed $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$, and any given $\tau$ and $\beta$, equation (12) gives, for any $\boldsymbol{b} \in \mathcal{B}$ with $\|\boldsymbol{b}\|_2 = 1$

$$\boldsymbol{b}^T \hat{H}_{r,\alpha,\beta} \boldsymbol{b} = \sum_{j=1}^{K} \frac{1}{\hat{\zeta}_j^2} (h_{r,\alpha,\beta}(\hat{\ell}_j^{\mathrm{e}}) - h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathrm{e}}))(\langle \boldsymbol{b}, \hat{\mathbf{p}}_j \rangle)^2 + h_{r,\alpha,\beta}(\hat{\ell}_0^{\mathrm{e}})\|\mathbf{b}\|^2$$

and from equations (10), (11), (16) and (17), the above reduces to

$$\begin{aligned} \boldsymbol{b}^T \hat{H}_{r,\alpha,\beta} \boldsymbol{b} &= \sum_{j=1}^{K} (h_{r,\alpha,\beta}(\ell_j) - h_{r,\alpha,\beta}(\ell_0))(\langle \mathbf{b}, \mathbf{p}_j \rangle)^2 + h_{r,\alpha,\beta}(\ell_0) + O_P(\sqrt{\log n/n}) \\ &= \boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b} + O_P(\sqrt{\log n/n}), \end{aligned} \tag{20}$$

uniformly over $\boldsymbol{b} \in \mathcal{B}$ consisting of $O(n^c)$ unit vectors, for any fixed $c > 0$. Next, since by assumption **A3**, $\tau$ and $\beta$ belong to compact subsets on which all the quantities in question are smooth functions with uniformly bounded Lischitz seminorm with respect $(\tau, \beta)$, by choosing appropriate grid of $(\tau, \beta)$ of size $O(n^{c'})$ for some $c' > 0$, we note that the expansion in (20) continue to hold uniformly in $(\tau, \beta)$, and hence we have $\sup_{\tau \in \mathbf{T}_0, \beta \in \mathbf{B}_0, \boldsymbol{b} \in \mathcal{B}} |\boldsymbol{b}^T \hat{H}_{r,\alpha,\beta} \boldsymbol{b} - \boldsymbol{b}^T H_{r,\alpha,\beta} \boldsymbol{b}| = O_P(\sqrt{\log n/n})$, thus proving the theorem.

## A.3 Proof of Theorem 1B

We only prove the result for fixed $(\tau, \beta)$ since the argument can be extended to compact subsets of $(\tau, \beta)$, under assumption **A3**, using an argument similar to that used in the proof of Theorem 1A.

Since the aggregated Bayes predictive rules involve quadratic forms of the form $\boldsymbol{b}^T G_{r,\alpha,\beta} \boldsymbol{b}$, we have the following cases of interest: $G_{0,1,0}, G_{1,0,\beta}$ and $G_{1,-1,\beta}$. In order to analyze the corresponding estimators of these quantities of interest, we introduce some notations. Let $\mathbf{C} = \mathbf{A}\mathbf{A}^T$, $\mathbf{Q} = [\mathbf{q}_1 : \cdots : \mathbf{q}_K]$, where $\mathbf{q}_j = \mathbf{A}\mathbf{p}_j$, and $\widetilde{\mathbf{Q}} = [\widetilde{\mathbf{q}}_1 : \cdots : \widetilde{\mathbf{q}}_K]$ where $\widetilde{\mathbf{q}}_j = \hat{\zeta}_j^{-1}\mathbf{A}\hat{\mathbf{p}}_j$. Then, for any $\beta \in \mathbb{R}^+$,

$$\begin{aligned}
\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T &= (\hat{\ell}^{\mathsf{e}}_0)^\beta \mathbf{A}\mathbf{A}^T + \sum_{j=1}^K \frac{(\hat{\ell}^{\mathsf{e}}_j)^\beta - (\hat{\ell}^{\mathsf{e}}_0)^\beta}{\hat{\zeta}_j^2}\mathbf{A}\hat{\mathbf{p}}_j\hat{\mathbf{p}}_j^T\mathbf{A}^T \\
&= (\hat{\ell}^{\mathsf{e}}_0)^\beta \mathbf{C} + \sum_{j=1}^K \left((\hat{\ell}^{\mathsf{e}}_j)^\beta - (\hat{\ell}^{\mathsf{e}}_0)^\beta\right)\widetilde{\mathbf{q}}_j\widetilde{\mathbf{q}}_j^T \;=\; (\hat{\ell}^{\mathsf{e}}_0)^\beta \left[\mathbf{C} + \widetilde{\mathbf{Q}}\left((\hat{\ell}^{\mathsf{e}}_0)^{-\beta}\hat{\Lambda}^\beta - \mathbf{I}_K\right)\widetilde{\mathbf{Q}}^T\right]
\end{aligned} \tag{21}$$

Setting $\beta = 1$, we observe that $\boldsymbol{b}^T\hat{G}_{0,1,0}\mathbf{b} = \hat{\ell}^{\mathsf{e}}_0\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{b} + \sum_{j=1}^K(\hat{\ell}^{\mathsf{e}}_j - \hat{\ell}^{\mathsf{e}}_0)\frac{1}{\hat{\zeta}_j^2}\left(\langle\boldsymbol{b}, \boldsymbol{A}\hat{\mathbf{p}}_j\rangle\right)^2$ which is $\boldsymbol{b}^T G_{0,1,0}\mathbf{b} + O_P(\sqrt{\log n/n})$ from equations (19) and (17). This proves the theorem when $\beta = 1$.

To prove the theorem for any $\beta \neq 1$, we make repeated use of the following basic formula for matrix inversion. Given a symmetric nonsingular $p \times p$ matrix $\mathbf{B}$, and a $p \times q$ matrix $\mathbf{D}$,

$$\left(\mathbf{B} + \mathbf{D}\mathbf{D}^T\right)^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{D}\left(\mathbf{I} + \mathbf{D}^T\mathbf{B}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}^T\mathbf{B}^{-1}. \tag{22}$$

Using (22) and (21), we have, with $\hat{\Delta}_\beta = (\hat{\ell}^{\mathsf{e}}_0)^{-\beta}\hat{\Lambda}^\beta - \mathbf{I}_K$,

$$(\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T)^{-1} = (\hat{\ell}^{\mathsf{e}}_0)^{-\beta}\mathbf{C}^{-1} - (\hat{\ell}^{\mathsf{e}}_0)^{-\beta}\mathbf{C}^{-1}\widetilde{\mathbf{Q}}\hat{\Delta}_\beta^{1/2}\left[\mathbf{I}_K + \hat{\Delta}_\beta^{1/2}\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1}\widetilde{\mathbf{Q}}\hat{\Delta}_\beta^{1/2}\right]^{-1}\hat{\Delta}_\beta^{1/2}\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1}$$

We must therefore analyze the behavior of $\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1}\widetilde{\mathbf{Q}}$. As a preliminary step, we observe that since $\mathbf{A}\mathbf{P}_{K,\perp}\boldsymbol{\varepsilon}_j \sim N(0, \mathbf{A}(\mathbf{I} - \mathbf{P}_K\mathbf{P}_K^T)\mathbf{A}^T)$, it follows that

$$\frac{1}{p}\|\mathbf{C}^{-1/2}\mathbf{A}\mathbf{P}_{K,\perp}\boldsymbol{\varepsilon}_j\|^2 = \frac{1}{p}\mathrm{trace}\left(\mathbf{C}^{-1}\mathbf{A}(\mathbf{I} - \mathbf{P}_K\mathbf{P}_K^T)\mathbf{A}^T\right) + O_P(p^{-1/2}) \tag{23}$$

which reduces to $1 - r_A/p + O_P(p^{-1/2})$ where $r_A = p^{-1}\mathrm{trace}\left(\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}\mathbf{P}_K\mathbf{P}_K^T\right)$ and

$$|r_A| \leq \|\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}\| \,\cdot\, \mathrm{rank}(\mathbf{P}_K\mathbf{P}_K^T) \leq K.$$

We will use equation (23) and the bound on $|r_A|$ to control $\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1}\widetilde{\mathbf{Q}}$. First note that using (16), (17) and, for any $1 \leq j, k \leq K$,

$$\begin{aligned}
\widetilde{\mathbf{q}}_j^T\mathbf{C}^{-1}\widetilde{\mathbf{q}}_k &= \mathbf{p}_j^T\mathbf{A}^T\mathbf{C}^{-1}\mathbf{A}\mathbf{p}_k(1 + O_P(n^{-1/2})) \\
&\quad + \frac{\sqrt{1-\zeta_j^2}\sqrt{1-\zeta_k^2}}{\zeta_j\zeta_k}\frac{\langle\mathbf{A}\mathbf{P}_{K,\perp}\boldsymbol{\varepsilon}_k, \mathbf{C}^{-1}\mathbf{A}\mathbf{P}_{K,\perp}\boldsymbol{\varepsilon}_j}{n-K}(1 + O_P(n^{-1/2}))
\end{aligned}$$

which, using equation (23), reduces to

$$\mathbf{p}_j^T \mathbf{A}^T \mathbf{C}^{-1} \mathbf{A} \mathbf{p}_k (1 + O_P(n^{-1/2})) + \frac{\sqrt{1-\zeta_j^2}\sqrt{1-\zeta_k^2}}{\zeta_j \zeta_k} \frac{p - r_A + O_P(p^{1/2})}{n-K} (1 + O_P(n^{-1/2}))$$

and finally to $\mathbf{q}_j^T \mathbf{C}^{-1} \mathbf{q}_k (1 + O_P(n^{-1/2})) + O_P(p/n)$ using the bound on $|r_A|$. Consequently, we have

$$\widetilde{\mathbf{Q}}^T \mathbf{C}^{-1} \widetilde{\mathbf{Q}} = \mathbf{Q}^T \mathbf{C}^{-1} \mathbf{Q} + O_P(n^{-1/2}) + O_P(p/n). \tag{24}$$

Now let $\Delta_\beta = \ell_0^{-\beta} \Lambda - \mathbf{I}_K$. Then $\|\hat{\Delta}_\beta - \Delta_\beta\| = O_P(n^{-1/2})$, and hence, by (24),

$$\hat{\mathbf{U}}_\beta^{-1} := \left[\mathbf{I}_K + \hat{\Delta}_\beta^{1/2} \widetilde{\mathbf{Q}}^T \mathbf{C}^{-1} \widetilde{\mathbf{Q}} \hat{\Delta}_\beta^{1/2}\right]^{-1} = \left[\mathbf{I}_K + \Delta_\beta^{1/2} \mathbf{Q}^T \mathbf{C}^{-1} \mathbf{Q} \Delta_\beta^{1/2}\right]^{-1} + \mathbf{R}_{\beta,n} = \mathbf{U}_\beta^{-1} + \mathbf{R}_{\beta,n}, \tag{25}$$

where $\|\mathbf{R}_{\beta,n}\| = O_P(n^{-1/2}) + O_P(p/n)$. Furthermore, $(\boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T)^{-1} + \tau^{-1}(\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T)^{-1}$ can be written as

$$\left((\hat{\ell}_0^{\mathsf{e}})^{-1} + \frac{1}{\tau}(\hat{\ell}_0^{\mathsf{e}})^{-\beta}\right)\mathbf{C}^{-1} - \mathbf{C}^{-1}\widetilde{\mathbf{Q}}\left[(\hat{\ell}_0^{\mathsf{e}})^{-1}\hat{\Delta}_1^{1/2}\hat{\mathbf{U}}_1\hat{\Delta}_1^{1/2} + \frac{1}{\tau}(\hat{\ell}_0^{\mathsf{e}})^{-\beta}\hat{\Delta}_\beta^{1/2}\hat{\mathbf{U}}_\beta\hat{\Delta}_\beta^{1/2}\right]\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1} ,$$

which by (25) is

$$\left((\hat{\ell}_0^{\mathsf{e}})^{-1} + \frac{1}{\tau}(\hat{\ell}_0^{\mathsf{e}})^{-\beta}\right)\mathbf{C}^{-1} - \mathbf{C}^{-1}\widetilde{\mathbf{Q}}\hat{\mathbf{V}}_\beta\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1} \tag{26}$$

where $\hat{\mathbf{V}}_\beta = \mathbf{V}_\beta + \check{\mathbf{R}}_{1,n} + \frac{1}{\tau}\check{\mathbf{R}}_{\beta,n}$ with $\mathbf{V}_\beta = \ell_0^{-1}\Delta_1^{1/2}\mathbf{U}_1\Delta_1^{1/2} + \frac{1}{\tau}\ell_0^{-\beta}\Delta_\beta^{1/2}\mathbf{U}_\beta\Delta_\beta^{1/2}$ and $\check{\mathbf{R}}_{\beta,n} = \hat{\ell}_0^{-\beta}\hat{\Delta}_\beta^{1/2}\mathbf{R}_{\beta,n}\hat{\Delta}_\beta^{1/2}$, so that $\|\check{\mathbf{R}}_{\beta,n}\| = O_P(n^{-1/2}) + O_P(p/n)$ for all $\beta$. Notice that $\mathbf{V}_\beta$ is positive definite, and hence $\hat{\mathbf{V}}_\beta$ is positive definite with probability tending to 1.

Define, for $x > 0$, $a_{\beta,\tau}(x) = x^{-1} + \tau^{-1}x^{-\beta}$. By (26), we can write $\left[(\boldsymbol{A}\hat{H}_{0,1,0}\boldsymbol{A}^T)^{-1} + \tau^{-1}(\boldsymbol{A}\hat{H}_{0,\beta,0}\boldsymbol{A}^T)^{-1}\right]^{-1}$ as

$$\frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})}\mathbf{C}\left[\mathbf{C} - \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})}\widetilde{\mathbf{Q}}\hat{\mathbf{V}}_\beta\widetilde{\mathbf{Q}}^T\right]^{-1}\mathbf{C} = \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})}\mathbf{C} + \frac{1}{(a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}}))^2}\widetilde{\mathbf{Q}}\left[\hat{\mathbf{V}}_\beta^{-1} - \frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})}\widetilde{\mathbf{Q}}^T\mathbf{C}^{-1}\widetilde{\mathbf{Q}}\right]^{-1}\widetilde{\mathbf{Q}}^T$$

and using $\hat{\mathbf{V}}_\beta = \mathbf{V}_\beta + \check{\mathbf{R}}_{1,n} + \frac{1}{\tau}\check{\mathbf{R}}_{\beta,n}$, we can re-write it as

$$\frac{1}{a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}})}\mathbf{C} + \frac{1}{(a_{\beta,\tau}(\hat{\ell}_0^{\mathsf{e}}))^2}\widetilde{\mathbf{Q}}\left[\mathbf{V}_\beta^{-1} - \frac{1}{a_{\beta,\tau}(\ell_0)}\mathbf{Q}^T\mathbf{C}^{-1}\mathbf{Q} + \mathbf{R}_{*,n}\right]^{-1}\widetilde{\mathbf{Q}}^T$$

where $\|\mathbf{R}_{*,n}\| = O_P(n^{-1/2}) + O_P(p/n)$. As a consequence, we have

$$\boldsymbol{b}^T \hat{G}_{1,0,\beta}\boldsymbol{b} = \boldsymbol{b}^T G_{1,0,\beta}\boldsymbol{b} + O_P(\sqrt{\log n/n}) + O_P(p/n)$$

uniformly over $\boldsymbol{b} \in \mathcal{B}$. An analogous calculation yields

$$\boldsymbol{b}^T \hat{G}_{1,-1,\beta}\boldsymbol{c} = \boldsymbol{b}^T G_{1,-1,\beta}\boldsymbol{c} + O_P(\sqrt{\log n/n}) + O_P(p/n)$$

uniformly over $\boldsymbol{b}, \boldsymbol{c} \in \mathcal{B}$.

# References

Aitchison, J. and I. R. Dunsmore (1976). Statistical prediction analysis. *Bulletin of the American Mathematical Society 82*(5), 683–688.

Baik, J. and J. W. Silverstein (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis 97*(6), 1382–1408.

Banerjee, T., G. Mukherjee, and W. Sun (2017). Adaptive sparse estimation with side information.

Benaych-Georges, F. and R. R. Nadakuditi (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis 111*, 120–135.

Binkiewicz, N., J. T. Vogelstein, and K. Rohe (2014). Covariate-assisted spectral clustering. *arXiv preprint arXiv:1411.2158*.

Bronnenberg, B. J., J.-P. H. Dubé, and M. Gentzkow (2012). The evolution of brand preferences: Evidence from consumer migration. *American Economic Review 102*(6), 2472–2508.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008). Database paper—the iri marketing data set. *Marketing science 27*(4), 745–748.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics*, 1685–1704.

Brown, L. D., G. Mukherjee, and A. Weinstein (2018). Empirical bayes estimates for a 2-way cross-classified additive model. *Annals of Statistics*.

Cai, T. T., Z. Ma, and Y. Wu (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics 41*(6), 3074–3110.

Cai, T. T., W. Sun, and W. Wang (2016). Cars: Covariate assisted ranking and screening for large-scale two-sample inference. *Technical Report*.

Cavrois, M., T. Banerjee, G. Mukherjee, N. Raman, R. Hussien, B. A. Rodriguez, J. Vasquez, M. H. Spitzer, N. H. Lazarus, J. J. Jones, et al. (2017). Mass cytometric analysis of hiv entry, replication, and remodeling in tissue cd4+ t cells. *Cell reports 20*(4), 984–998.

Coibion, O., Y. Gorodnichenko, and G. H. Hong (2015). The cyclicality of sales, regular and effective prices: Business cycle and policy implications. *American Economic Review 105*(3), 993–1029.

Dicker, L. H. and S. D. Zhao (2016). High-dimensional classification via nonparametric empirical bayes and maximum likelihood inference. *Biometrika 103*(1), 21–34.

Dobriban, E., W. Leeb, and A. Singer (2017). Optimal prediction in the linearly transformed spiked model. *arXiv preprint arXiv:1709.03393*.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

Efron, B. and T. Hastie (2016). *Computer age statistical inference*, Volume 5. Cambridge University Press.

El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics 36*(6), 2757–2790.

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Fan, J., W. Wang, and Y. Zhong (2016). Robust covariance estimation for approximate factor models. *arXiv preprint arXiv:1602.00719*.

Fourdrinier, D., W. E. Strawderman, and M. T. Wells (2017). *Shrinkage Estimation*. Springer.

Geisser, S. (1993). Predictive inference. monographs on statistics & applied probability. *CRC, London*.

George, E. I., F. Liang, and X. Xu (2006). Improved minimax predictive densities under kullback–leibler loss. *The Annals of Statistics 34*(1), 78–91.

George, E. I. and X. Xu (2008). Predictive density estimation for multiple regression. *Econometric Theory 24*(2), 528–544.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models*.

Granger, C. W. (1969). Prediction with a generalized cost of error function. *Journal of the Operational Research Society 20*(2), 199–207.

Green, L. V., S. Savin, and N. Savva (2013). "nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science 59*(10), 2237–2256.

Greenshtein, E. and J. Park (2009). Application of non parametric empirical bayes estimation to high dimensional classification. *Journal of Machine Learning Research 10*(Jul), 1687–1704.

Greenshtein, E. and Y. Ritov (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium*, Volume 57, pp. 266–275.

Harvey, C. R., Y. Liu, and H. Zhu (2016). ... and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Johnstone, I. M. and A. Y. Lu (2012). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*.

Johnstone, I. M. and D. Paul (2018). Pca in high dimensions: An orientation. *Proceedings of the IEEE* (99).

Johnstone, I. M. and D. M. Titterington (2009). Statistical challenges of high-dimensional data.

Karoui, N. E., A. E. Lim, and G.-Y. Vahn (2011). Estimation error reduction in portfolio optimization with conditional value-at-risk. Technical report.

Ke, T., J. Jin, and J. Fan (2014). Covariance assisted screening and estimation. *Annals of statistics 42*(6), 2202.

Khamaru, K. and R. Mazumder (2018). Computation of the maximum likelihood estimator in low-rank factor analysis. *arXiv preprint arXiv:1801.05935*.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Komaki, F. (2015). Asymptotic properties of bayesian predictive densities when the distributions of data and target variables are different. *Bayesian Analysis 10*(1), 31–51.

Kong, X., Z. Liu, P. Zhao, and W. Zhou (2017). Sure estimates under dependence and heteroscedasticity. *Journal of Multivariate Analysis 161*, 1–11.

Kou, S. and J. J. Yang (2015). Optimal shrinkage estimation in heteroscedastic hierarchical linear models. *arXiv preprint arXiv:1503.06262*.

Kozak, S., S. Nagel, and S. Santosh (2017). Shrinking the cross section. *SSRN*.

Kritchman, S. and B. Nadler (2008). Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems 94*(1), 19–32.

Kritchman, S. and B. Nadler (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing 57*(10), 3930–3941.

Levi, R., G. Perakis, and J. Uichanco (2015). The data-driven newsvendor problem: new bounds and insights. *Operations Research 63*(6), 1294–1306.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801.

Mukherjee, G., L. D. Brown, and P. Rusmevichientong (2015). Efficient empirical bayes prediction under check loss using asymptotic risk estimates. *arXiv preprint arXiv:1511.00028*.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

Onatski, A. (2015). Asymptotic analysis of the squared estimation error in misspecified factor models. *Journal of Econometrics 186*(2), 388–406.

Onatski, A., M. J. Moreira, and M. Hallin (2014). Signal detection in high dimension: The multispiked case. *The Annals of Statistics 42*(1), 225–254.

Owen, A. B. and J. Wang (2016). Bi-cross-validation for factor analysis. *Statistical Science 31*(1), 119–139.

Passemier, D., Z. Li, and J. Yao (2015). On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Passemier, D., Z. Li, and J. Yao (2017). On estimation of the noise variance in high dimensional probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 51–67.

Passemier, D. and J.-F. Yao (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices: Theory and Applications 1*(01), 1150002.

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.

Patton, A. J. and A. Timmermann (2007). Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics 140*(2), 884–918.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 1617–1642.

Paul, D. and A. Aue (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference 150*, 1–29.

Press, S. J. (2009). *Subjective and objective Bayesian statistics: principles, models, and applications*, Volume 590. John Wiley & Sons.

Robbins, H. (1985). The empirical bayes approach to statistical decision problems. In *Herbert Robbins Selected Papers*, pp. 49–68. Springer.

Rudin, C. and G.-Y. Vahn (2014). The big data newsvendor: Practical insights from machine learning.

Sen, N., G. Mukherjee, A. Sen, S. C. Bendall, P. Sung, G. P. Nolan, and A. M. Arvin (2014). Single-cell mass cytometry analysis of human tonsil t cell remodeling by varicella zoster virus. *Cell reports 8*(2), 633–645.

Sun, W. and T. Cai (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 393–424.

Tan, Z. (2015). Improved minimax estimation of a multivariate normal mean under heteroscedasticity. *Bernoulli 21*(1), 574–603.

Varian, H. R. (1975). A bayesian approach to real estate assessment. *Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage*, 195–208.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2015). Group-linear empirical bayes estimates for a heteroscedastic normal mean. *arXiv preprint arXiv:1503.08503*.

Xie, X., S. Kou, and L. D. Brown (2012). Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association 107*(500), 1465–1479.

Xie, X., S. C. Kou, and L. Brown (2016). Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *Annals of statistics 44*(2), 564.

Yano, K. and F. Komaki. Information criteria for prediction when distributions of data and target variables are different. *Statistica Sinica*.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association 81*(394), 446–451.

Zellner, A. and M. S. Geisel (1968). *Sensitivity of control to uncertainty and form of the criterion function*. U. of Chicago.

Zhang, C.-H. (2003, 04). Compound decision theory and empirical bayes methods: invited paper. *Ann. Statist. 31*(2), 379–390.

# Supplementary Material for "Improved Shrinkage Prediction under a Spiked Covariance Structure"

The supplement contains the proofs of Lemmata 1, 2, 3 and 4 as well as that of Theorems 2A and 2B.

## 1 Proof of Lemma 1

First note that under the hierarchical models (1) and (4), the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{AX}$ is $N(\boldsymbol{A\eta}_0 + G_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\eta}_0), G_{1,0,\beta})$. To prove this Lemma, we first fix a few notations. For coordinate $i$, let $v_i, v_{\psi i}$ and $v_{fi}$ denote the $i^{th}$ diagonal element of $\check{\boldsymbol{\Sigma}}_1$, $\check{\boldsymbol{\Sigma}}_\beta$ and $m_0^{-1}\check{\boldsymbol{\Sigma}}_1$ respectively. The minimizer of the univariate Bayes risk $B_i(\tau,\beta)$ is given by

$$\hat{q}_i = \arg\min_q \int L_i(\psi_i, q_i)\pi(\psi_i|(\boldsymbol{AX})_i)$$

where the posterior distribution $\pi(\psi_i|(\boldsymbol{AX})_i) \sim N(\gamma_i, \omega_i)$ where $\gamma_i = \delta_i(\boldsymbol{AX})_i + (1-\delta_i)(\boldsymbol{A\eta}_0)_i$, $\delta_i = v_{\psi i}/(v_{\psi i}+v_i)$ and $\omega_i = (v_{\psi i}^{-1} + v_i^{-1})^{-1}$. We prove the Lemma for the generalized absolute loss and the linex loss functions. The univariate Bayes predictive rules for the other losses considered in this paper will follow from similar arguments.

For the linex loss function, note that $L_i(\psi_i, q_i) = \mathbb{E}_{V_i}\mathcal{L}_i(V_i, q_i)$ where $\mathcal{L}_i(V_i, q_i)$ is the linex loss for coordinate $i$ from equation (6). Since $V_i \sim N(\psi_i, v_{fi})$,

$$\mathbb{E}_{V_i}\mathcal{L}_i(V_i, q_i) = b_i\Big[\exp\{a_i(q_i-\psi_i) + (a_i^2/2)v_{fi}\} - a_i(q_i-\psi_i) - 1\Big]$$

Furthermore, $\mathbb{E}_{\psi_i|(\boldsymbol{AX})_i}L_i(\psi_i, q_i) = b_i\Big[\exp\{a_i(q_i-\gamma_i) + (a_i^2/2)(v_{fi}+\omega_i)\} - a_i(q_i-\gamma_i) - 1\Big]$ is convex in $q_i$. Differentiating the above posterior expectation with respect to $q_i$, we get

$$\hat{q}_i = \delta_i(\boldsymbol{AX})_i + (1-\delta_i)(\boldsymbol{A\eta}_0)_i - \frac{a_i}{2}(v_{fi}+\omega_i)$$

which completes the proof.

For the generalized absolute loss function in equation (5), note that

$$\mathbb{E}_{V_i}\mathcal{L}_i(V_i, q_i) = b_i(\psi_i - q_i) + (b_i + h_i)\mathbb{E}(q_i - \psi_i - Z)^+$$

where $Z$ is a standard normal random variable. Furthermore, direct calculation yields $\mathbb{E}(q_i - \psi_i - Z)^+ = (q_i - \psi_i)\Phi(q_i - \psi_i) + \phi(q_i - \psi_i)$. The Bayes predictive rule then follows from Lemma 2.1 and 2.2 of Mukherjee et al. (2015).

# 2 Proof of Lemma 2

We prove this lemma for the generalized absolute loss function in equation (5). For any $i$ and fixed $(\tau, \beta)$, it follows from Theorem 1A,

$$\left|\hat{q}_i^{\mathsf{approx}}(\boldsymbol{X}|\boldsymbol{S}, \tau, \beta) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\boldsymbol{\Sigma}, \tau, \beta)\right| \leq \left|\boldsymbol{e}_i^T(\hat{H}_{1,-1,\beta} - H_{1,-1,\beta})(\boldsymbol{X} - \boldsymbol{\eta}_0)\right| + O_p\Big(\sqrt{\frac{\log n}{n}}\Big) \quad (1)$$

The first term on the right of the inequality above is an asymmetric quadratic form and can be written as a difference of two symmetric quadratic forms as follows

$$\frac{\|\boldsymbol{X} - \boldsymbol{\eta}_0\|_2}{4}\Big[\Big|(\boldsymbol{a} + \boldsymbol{e}_i)^T(\hat{H}_{1,-1,\beta} - H_{1,-1,\beta})(\boldsymbol{a} + \boldsymbol{e}_i)\Big| - \Big|(\boldsymbol{a} - \boldsymbol{e}_i)^T(\hat{H}_{1,-1,\beta} - H_{1,-1,\beta})(\boldsymbol{a} - \boldsymbol{e}_i)\Big|\Big]$$

where $\boldsymbol{a} = (\boldsymbol{X} - \boldsymbol{\eta}_0)/\|\boldsymbol{X} - \boldsymbol{\eta}_0\|_2$.

Re-apply Theorem 1A separately to these two symmetric quadratic forms and note that the above is bounded by $O_p(\sqrt{\log n/n})(\|\boldsymbol{X} - \boldsymbol{\eta}_0\|_2)(\|\boldsymbol{a} + \boldsymbol{e}_i\|_2^2 + \|\boldsymbol{a} - \boldsymbol{e}_i\|_2^2)/4$, from which the result follows.

# 3 Proofs of Lemma 3 and Lemma 4

To prove these lemmas, we use the following result.

**Lemma A.** *Under assumptions **A1** and **A2**, uniformly in $\boldsymbol{b} \in \mathcal{B}$ such that $\mathcal{B} = O(n^c)$ for any fixed $c > 0$, with $\|\boldsymbol{b}\|_2 = 1$, and for all $(r, \alpha) \in \{-1, 0, 1\} \times \mathbb{R}$, we have as $n \to \infty$,*

$$\sup_{\boldsymbol{b}\in\mathcal{B}}\Big|\boldsymbol{b}^T\hat{H}_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})\hat{H}_{1,-1,\beta}\boldsymbol{b} - \boldsymbol{b}^T H_{1,-1,\beta}\mathcal{J}(\boldsymbol{\Sigma})H_{1,-1,\beta}\boldsymbol{b} - j(\ell_0)\sum_{j=1}^{K}(h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2(\langle \boldsymbol{b}, \mathbf{p}_j\rangle)^2\Big| = O_P(\sqrt{\log n/n})\,.$$

**Proof of Lemma A.** Let us first define the following quantities: $\Delta_j(h_1) = h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0)$, $\Delta_j = \mathcal{J}(\ell_j) - \mathcal{J}(\ell_0)$ and $\hat{\Delta}_j(h_1) = h_{1,-1,\beta}(\hat{\ell}^{\mathsf{e}}_j) - h_{1,-1,\beta}(\hat{\ell}^{\mathsf{e}}_0)$ where $h_{r,\alpha,\beta}$ is the scalar version of $H_{r,\alpha,\beta}$ and $\mathcal{J}(x) = x + \tau x^{\beta}$ being the scalar version of $\mathcal{J}(\mathbf{\Sigma})$. For any $\boldsymbol{b} \in \mathcal{B}$ with $\|\boldsymbol{b}\|_2 = 1$, expand $\boldsymbol{b}^T \hat{H}_{1,-1,\beta}\mathcal{J}(\mathbf{\Sigma})\hat{H}^T_{1,-1,\beta}\boldsymbol{b}$ as

$$
\begin{aligned}
&\sum_{j=1}^{K}\sum_{j'=1}^{K}\sum_{k=1}^{K} \frac{\hat{\Delta}_j(h)\hat{\Delta}_{j'}(h)}{\hat{\zeta}_j^2\hat{\zeta}_{j'}^2}\Delta_k\langle \boldsymbol{b}, \hat{\mathbf{p}}_j\rangle\langle \mathbf{b}, \hat{\mathbf{p}}_{j'}\rangle\langle \mathbf{p}_k, \hat{\mathbf{p}}_j\rangle\langle \mathbf{p}_k, \hat{\mathbf{p}}_{j'}\rangle \\
&+\ \mathcal{J}(\ell_0)\sum_{j=1}^{K}\sum_{j'=1}^{K} \frac{\hat{\Delta}_j(h)\hat{\Delta}_{j'}(h)}{\hat{\zeta}_j^2\hat{\zeta}_{j'}^2}\langle \mathbf{b}, \hat{\mathbf{p}}_j\rangle\langle \mathbf{b}, \hat{\mathbf{p}}_{j'}\rangle\langle \hat{\mathbf{p}}_{j'}, \hat{\mathbf{p}}_j\rangle \\
&+\ 2h(\hat{\ell}^{\mathsf{e}}_0)\sum_{j=1}^{K}\sum_{k=1}^{K}\frac{\hat{\Delta}_j(h)}{\hat{\zeta}_j^2}\Delta_k\langle \mathbf{b}, \hat{\mathbf{p}}_j\rangle\langle \mathbf{b}, \mathbf{p}_k\rangle\langle \mathbf{p}_k, \hat{\mathbf{p}}_j\rangle + 2h(\hat{\ell}^{\mathsf{e}}_0)\mathcal{J}(\ell_0)\sum_{j=1}^{K}\frac{\hat{\Delta}_j(h)}{\hat{\zeta}_j^2}(\langle \mathbf{b}, \hat{\mathbf{p}}_j\rangle)^2 \\
&+\ (h(\hat{\ell}^{\mathsf{e}}_0))^2\sum_{k=1}^{K}\Delta_k(\langle \mathbf{b}, \mathbf{p}_k\rangle)^2 + (h(\hat{\ell}^{\mathsf{e}}_0))^2\mathcal{J}(\ell_0)\|\mathbf{b}\|^2
\end{aligned}
$$

Then, using equation (9), it can be verified that above asymptotically equals

$$
\begin{aligned}
&\boldsymbol{b}^T\Big(\sum_{j=1}^{K}\Delta_j(h_1)\mathbf{p}_j\mathbf{p}_j^T + h_{1,-1,\beta}(\ell_0)\boldsymbol{I}\Big)\Big(\sum_{j=1}^{K}\Delta_j\mathbf{p}_j\mathbf{p}_j^T + h_{1,-1,\beta}(\ell_0)\boldsymbol{I}\Big) \\
&\quad\Big(\sum_{j=1}^{K}\Delta_j(h_1)\mathbf{p}_j\mathbf{p}_j^T + h_{1,-1,\beta}(\ell_0)\boldsymbol{I}\Big)\boldsymbol{b} +\ \mathcal{J}(\ell_0)\sum_{j=1}^{K}\Big(\Delta_j(h_1)\Big)^2(\langle \boldsymbol{b}, \mathbf{p}_j\rangle)^2 + O_P(\sqrt{\log n/n})
\end{aligned}
$$

where the $O_P$ term is uniform in $\boldsymbol{b} \in \mathcal{B}$ consisting of $O(n^c)$ unit vectors. Finally, using the definitions of $\Delta_j(h_1)$, $\Delta_j$ and arguments similar that used in proving Theorem 1A, the result follows.

Next, we prove the three statements of the lemma.

**Proof of Lemma 3, statement (a)** - first note that $\mathbb{E}\Big\{\Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i, \tau, \beta) - q_i^{\mathsf{Bayes}}(\boldsymbol{X}|\mathbf{\Sigma}, \tau, \beta)\Big)^2\Big\}$ can be decomposed as

$$
\mathbb{E}^2\Big\{\Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i) - q_i^{\mathsf{Bayes}}(\boldsymbol{X})\Big)\Big\} + \mathsf{Var}\Big\{\Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i) - q_i^{\mathsf{Bayes}}(\boldsymbol{X})\Big)\Big\}
$$

where the first term represents bias squared and the second term is the variance. Now consider, for example, the generalized absolute loss function of equation (5). Under this loss, the bias with respect to the marginal distribution of $\boldsymbol{X}$ is

$$
\Phi^{-1}(\tilde{b}_i)\Big[\Big(\mathbf{e}_i^T\hat{H}_{1,0,\beta}\mathbf{e}_i + m_0^{-1}\mathbf{e}_i^T\hat{H}_{0,1,0}\mathbf{e}_i\Big)^{1/2} - \Big(\mathbf{e}_i^T H_{1,0,\beta}\mathbf{e}_i + m_0^{-1}\mathbf{e}_i^T H_{0,1,0}\mathbf{e}_i\Big)^{1/2}\Big]
$$

which, by Theorem 1A is $O_P(\sqrt{\log n/n})$. Now the variance term is equal to

$$f_i^2 \boldsymbol{e}_i^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) \hat{H}_{1,-1,\beta}^T \boldsymbol{e}_i - 2 f_i \boldsymbol{e}_i^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i + \boldsymbol{e}_i^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i \ ,$$

which is a quadratic with respect to $f_i$ and is minimized at

$$f_i^{\mathsf{OR}} = \frac{\boldsymbol{e}_i^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i}{\boldsymbol{e}_i^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) \hat{H}_{1,-1,\beta}^T \boldsymbol{e}_i} \ .$$

The numerator in the above expression is an asymmetric quadratic form in $\hat{H}_{r,\alpha,\beta}$ and by Theorem 1A it equals $\boldsymbol{e}_i^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i + (\boldsymbol{e}_i^T H_{1,-1,\beta} \mathcal{J}^2(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i)^{1/2} O_p(\sqrt{\log n/n})$. By Lemma A, the denominator is

$$\boldsymbol{e}_i^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta} \boldsymbol{e}_i + j(\ell_0) \sum_{j=1}^{K} (h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2 (\langle \boldsymbol{e}_i, \mathbf{p}_j \rangle)^2 + O_p(\sqrt{\log n/n})$$

which, for fixed $\tau > 0, \beta \geq 0$, is non-trivial since $\ell_j > \ell_0 > 0$ for all $j = 1, \ldots, K$. Thus, the ratio asymptotically equals

$$\frac{\boldsymbol{e}_i^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i}{\boldsymbol{e}_i^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}^T \boldsymbol{e}_i + j(\ell_0) \sum_{j=1}^{K} (h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2 (\langle \boldsymbol{e}_i, \mathbf{p}_j \rangle)^2} + O_p\left(\sqrt{\frac{\log n}{n}}\right).$$

The second term in the denominator is at least as big as

$$j(\ell_0)(h_{1,-1,\beta}(\ell_K) - h_{1,-1,\beta}(\ell_0))^2 \|\mathbf{P}_K \boldsymbol{e}_i\|^2.$$

Finally, note that $U(\boldsymbol{\Sigma}) = H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta}$, from which the result follows.

**Proof of Lemma 3, statement (b)** - from Lemma A, $\boldsymbol{b}^T \hat{H}_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) \hat{H}_{1,-1,\beta}^T \boldsymbol{b}$ asymptotically equals

$$\boldsymbol{b}^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta} \boldsymbol{b} + j(\ell_0) \sum_{j=1}^{K} (h_{1,-1,\beta}(\ell_j) - h_{1,-1,\beta}(\ell_0))^2 (\langle \boldsymbol{b}, \mathbf{p}_j \rangle)^2 + O_p(\sqrt{\log n/n})$$

which is strictly bigger than $\boldsymbol{b}^T H_{1,-1,\beta} \mathcal{J}(\boldsymbol{\Sigma}) H_{1,-1,\beta} \boldsymbol{b} + O_P(\sqrt{\log n/n})$ for any fixed $\tau > 0, \beta > 0$ and from this the proof immediately follows.

**Proof of Lemma 3, statement (c)** - this follows directly from statements (a) and (b). For any coordinate $i$, by definition of $f_i^{\mathsf{OR}}$ in statement (a),

$$\mathbb{E}\Big[\Big(q_i^{\mathsf{approx}}(\boldsymbol{X}|\boldsymbol{S}) - q_i^{\mathsf{Bayes}}(\boldsymbol{X})\Big)^2\Big] \geq \mathbb{E}\Big[\Big(q_i^{\mathsf{cs}}(\boldsymbol{X}|\boldsymbol{S}, f_i^{\mathsf{OR}}) - q_i^{\mathsf{Bayes}}(\boldsymbol{X})\Big)^2\Big]$$

while statement (b) implies that the above inequality holds for all $i$ and for any fixed $\tau > 0$ and $\beta > 0$.

**Proof of Lemma 4** - The proof of this Lemma follows directly using Theorem 1A for the numerator of $\hat{f}_i^{\mathsf{prop}}$ and Lemma A and equations (8), (9) for the denominator. Similar arguments using Theorem 1B, Lemma A and equations (8), (9) prove the result for $\hat{f}_i^{\mathsf{prop}}$ in definition 4.

# 4 Proof of Theorems 2A and 2B

We will first prove Theorem 2B for the generalized absolute loss function in equation (5).

For any $i$ and fixed $\tau > 0$, $\beta > 0$, we have ,

$$\left|\hat{q}_i^{\mathsf{caspr}}(\hat{f}_i^{\mathsf{prop}}) - \hat{q}_i^{\mathsf{cs}}(f_i^{\mathsf{OR}})\right| = \left|\hat{f}_i^{\mathsf{prop}} - f_i^{\mathsf{OR}}\right|\left|\boldsymbol{e}_i^T \hat{H}_{1,-1,\beta}\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{\eta}_0)\right|$$

which can be upper bounded by $\left|1 - \hat{f}_i^{\mathsf{prop}}/f_i^{\mathsf{OR}}\right|\left[\left|\hat{q}_i^{\mathsf{cs}}(f_i^{\mathsf{OR}}) - \boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0\right| + \left|\Phi^{-1}(\tilde{b}_i)\right|\left(\mathbf{e}_i^T\hat{G}_{1,0,\beta}\mathbf{e}_i + m_0^{-1}\mathbf{e}_i^T\hat{G}_{0,1,0}\mathbf{e}_i\right)^{1/2}\right]$. Now using Theorem 1B and Lemma 4, $\|\hat{\boldsymbol{q}}^{\mathsf{caspr}}(\boldsymbol{AX}|\hat{\boldsymbol{f}}^{\mathsf{prop}}) - \hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{AX}|\boldsymbol{f}^{\mathsf{OR}})\|_2^2$ is upper bounded by

$$\frac{2}{\left(f_{\mathsf{inf}}^{\mathsf{OR}}\right)^2}\left[\|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{AX}|\boldsymbol{f}^{\mathsf{OR}}) - \boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0\|_2^2 + \left\{\Phi^{-1}(\tilde{b}_i)\right\}^2\left(\mathbf{e}_i^T G_{1,0,\beta}\mathbf{e}_i + m_0^{-1}\mathbf{e}_i^T G_{0,1,0}\mathbf{e}_i + c_n\right)\right] O_p\left(\frac{\log n}{n}\right)$$

where $f_{\mathsf{inf}}^{\mathsf{OR}} := \inf_{1 \le i \le n} f_i^{\mathsf{OR}} > 0$ and $c_n = O_p\left\{\max\left(\frac{p}{n}, \sqrt{\frac{\log n}{n}}\right)\right\}$. The proof then follows by noting that $\left|\left|\hat{\boldsymbol{q}}^{\mathsf{cs}}(\boldsymbol{AX}|\boldsymbol{f}^{\mathsf{OR}}) - \boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{\eta}_0\right|\right|_2^2 > 0$ since $\ell_0 > 0$. The proof of Theorem 2A follows using similar arguments with Theorem 1A and Lemma 4.

# References

Mukherjee, G., L. D. Brown, and P. Rusmevichientong (2015). Efficient empirical bayes prediction under check loss using asymptotic risk estimates. *arXiv preprint arXiv:1511.00028*.